ABSTRACT

Bayesian Evaluation and Adaptive Trial Design for Surrogate Time-to-Event Endpoints in Clinical Trials

Lindsay A. Renfro, Ph.D.

Chairperson: Bradley P. Carlin, Ph.D.

Surrogate endpoints are desirable in clinical trials when primary endpoints are costly to obtain, difficult to measure, or require lengthy follow-up to observe. Despite legitimate concerns, evaluation of potentially beneficial treatments in some settings remains impossible or implausible without the use of surrogates. Furthermore, strong evidence based on a collection of trials, rather than a relationship observed within a single trial, is required to validate a surrogate endpoint for future primary use. We present a Bayesian approach to evaluating surrogacy using patient data from multiple trials with time-to-event endpoints that accounts for estimation error of treatment effects and offers greater computational stability than existing methods.

Once a surrogate endpoint has been deemed valid for use in a future trial, a healthy skepticism should remain regarding its ability to reflect the true treatment effect that would have been observed on the primary endpoint. Despite the surrogate's intended role, few (if any) efforts have been made to formalize existing knowledge and uncertainty in the design of such a trial. We propose a Bayesian adaptive design that uses the validated surrogate as the primary endpoint, while acknowledging that this endpoint is really a surrogate, and perhaps only a recently-validated one. At prospectively-defined checkpoints, we assess the performance of

the surrogate and decide whether to continue its use or switch consideration to the primary endpoint. Furthermore, our design incorporates other favorable aspects of Bayesian adaptive trials, including the ability to stop a trial early for treatment efficacy, inferiority, or trial futility.

Flowgraphs are useful for modeling diseases that are well-described by multi-state models, but for which Markov assumptions are inadequate and returns to previous states are possible. Furthermore, censoring and covariates may influence the distribution of waiting times between any two states, and to a differing degree for separate transitions within the same system. We discuss the construction and advantages of flowgraph models when used to describe cancer progression within two clinical trials, where our goal is improved modeling of treatment effects and prediction of patient outcomes for the purpose of more realistic surrogacy evaluation.

Bayesian Evaluation and Adaptive Trial Design for Surrogate Time-to-Event Endpoints in Clinical Trials

by

Lindsay A. Renfro, B.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

_____

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

_____

Bradley P. Carlin, Ph.D., Chairperson

_____

Daniel J. Sargent, Ph.D.

_____

Michael E. Sherr, Ph.D.

_____

James D. Stamey, Ph.D.

_____

Jack D. Tubbs, Ph.D.

_____

Dean M. Young, Ph.D.

Accepted by the Graduate School
May 2011

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES viii

ACKNOWLEDGMENTS ix

DEDICATION xi

1 Introduction 1

1.1 Bayesian Evaluation of Surrogate Time-to-Event Endpoints 1

1.2 Bayesian Adaptive Trial for a Newly Validated Surrogate 1

1.3 Flowgraph Modeling for Surrogacy Evaluation 2

2 Bayesian Adjusted R^2 for the Meta-Analytic Evaluation of Surrogate Time-to-Event Endpoints in Clinical Trials 4

2.1 Background 4

2.2 Two-Stage Model for Trial-Level Surrogacy 6

2.2.1 Review of Classical Approach 6

2.2.2 Issues with Maximum Likelihood Estimation 10

2.2.3 Bayesian Model for Second Stage 11

2.3 Simulation Study 13

2.3.1 Data Generation 13

2.3.2 Estimation of Two-Stage Model 16

2.3.3 Results 17

2.4 Application to Adjuvant Trials in Colon Cancer 24

2.4.1 ACCENT Data 24
2.4.2 Analysis of TTR Surrogacy for OS 26
2.4.3 Results: Unadjusted R^2 27
2.4.4 Results: Bayesian Adjusted R^2 29
2.5 Extension of the Bayesian Adjusted Model 30

3 Bayesian Adaptive Trial Design for a Newly Validated Surrogate Endpoint 34
3.1 Introduction 34
3.2 Bayesian Model for Historical Trials 36
3.3 Bayesian Adaptive Design 38
3.3.1 Checking Surrogacy 42
3.4 Simulation Study 44
3.4.1 Data Generation and Settings 46
3.4.2 Simulation Results 49
3.5 Example: Adaptive Monitoring of ACCENT Trials 54

4 Flowgraph Modeling of Disease Progression for Evaluating Surrogate Endpoints 59
4.1 Introduction 59
4.2 Review of Flowgraph Models and Methods 62
4.2.1 Model Construction and Transmittances 62
4.2.2 Solving Flowgraphs and Mason's Rule 64
4.2.3 Saddlepoint Approximation and Bayesian Estimation 66
4.2.4 Extension: Inclusion of Covariates 70
4.3 Flowgraph Models for Colon Cancer Trials 70
4.3.1 Trial N0147: Reversible Illness-Death Model 73
4.3.2 Trial N9741: Illness-Death Model with a Possibly Prolonged Initial State 75

4.4 Application to Surrogacy Evaluation 76

5 Conclusion 78

5.1 Bayesian Evaluation of Surrogate Time-to-Event Endpoints 78

5.2 Bayesian Adaptive Trial for a Newly Validated Surrogate 81

5.3 Flowgraph Modeling for Surrogacy Evaluation 82

BIBLIOGRAPHY 84

LIST OF FIGURES

2.1 Simulation results for trial-level surrogacy 21

2.2 Simulation results, continued 22

2.3 ACCENT treatment effects on S and T by trial 26

2.4 ACCENT results for possible model extensions 31

2.5 ACCENT results for chosen extended model 32

3.1 Treatment effect scatterplots for simulated trials and ACCENT trials shown with adaptive design conclusions 45

4.1 Simple flowgraph model with parallel branches and absorbing states . . . 63

4.2 Flowgraph models for colon cancer trials N0147 and N9741 72

LIST OF TABLES

2.1 Simulation settings for trial-level surrogacy evaluation 16
2.2 Simulation results 19
2.3 Simulation results, continued 20
2.4 ACCENT results for unadjusted and adjusted trial-level surrogacy 28
2.5 ACCENT results for extended Bayesian model 33

3.1 Simulation scenarios for good or poor surrogate endpoint in a new trial . 49
3.2 Simulation results for good surrogate endpoint 50
3.3 Simulation results for good surrogate endpoint, continued 51
3.4 Simulation results for poor surrogate endpoint 52
3.5 Simulation results for poor surrogate endpoint, continued 53
3.6 Conclusions for ACCENT trials under adaptive design 57

ACKNOWLEDGMENTS

I must start by thanking Dr. Mark Hamner, as he is solely responsible for sparking my interest in statistics and the first person who trained me to think like a statistician. The subsequent years he invested in my development as a future Ph.D. student in this area, and a Bayesian statistician in particular, will continue to pay off throughout my career. I owe much to the Rice University Statistics faculty who nurtured me as a first year student: Dr. David Scott, Dr. George Terrell, Dr. Marek Kimmel, and Dr. Kenneth Hess. I am also indebted to the Baylor Statistics faculty members who challenged me daily for the next several years, including Dr. James Stamey, who provided me with an important and thorough foundation in mathematical statistics; Dr. Dennis Johnston, who was always encouraging and willing to consider my questions carefully; Dr. Jane Harvill, who guided me through a course I found to be less intuitive and provided personal guidance as well; Dr. Tom Bratcher, who saved his most difficult and tedious theoretical lecture for my birthday and expressed his disappointment when I scored a 90.5 on one of his exams; Dr. Dean Young, who thoroughly critiqued my mathematical writing until the day of my defense and introduced me to the journal submission process; and Dr. Jack Tubbs, who navigated my unusual course through Baylor's Ph.D. program with professionalism and assurance.

I want to thank my first graduate mentor, Dr. John Seaman, Jr., for being particularly instrumental in my development as a Bayesian and critical thinker. His beautifully crafted lectures and notes provide an incomparable reference to what the combination of thorough investigation, hard work, and attention to detail can produce. My primary graduate mentor, Dr. Bradley Carlin, is also worthy of tremendous thanks for advising me with sustained enthusiasm and clear direction from

approximately 1,000 miles away. Throughout the course of nearly two years, he provided direct and valuable feedback to my writing and ideas, balanced by unmatched insight regarding professional development, leadership, and nuances of the Bayesian and biostatistics world communities. Dozens of phone calls, hundreds of emails, and dissertation meetings organized in three countries kept Dr. Carlin constantly involved not only in my work, but also in my personal growth. I am very grateful to my past and future Mayo Clinic mentor, Dr. Daniel Sargent, for selecting me (seemingly at random) from a pile of summer internship applicants. This internship challenged me in many dimensions and provided the inspiration for all three primary chapters of this dissertation. I must also thank Dr. Sargent for staying actively involved with my research long after my employment had officially ended, and for giving me the opportunity to return to Mayo in order to continue pursuing these fascinating problems.

Most importantly, I must thank my immensely wonderful parents, Dave and Paulett Renfro, for supporting all of my wild ideas–from becoming an artist to becoming a statistician–with unending love and support. Like much of my life, obtaining this degree was an indirect path with many hills and valleys, but my parents never wavered in their enthusiasm to join me for the ride. I am also grateful for the steady patience and encouragement given by my very significant other, Dr. Lyle McKinney, who first collaborated with me on institutional research, and now joins me in our greatest joint effort–our first child, a son. And to my son, my tiny co-author, endless love and a very heartfelt thank you for bringing this academic journey to a close as smoothly and sweetly as possible.

DEDICATION

For Anna

CHAPTER ONE

Introduction

1.1 Bayesian Evaluation of Surrogate Time-to-Event Endpoints

Burzykowski et al. (2001) introduced a two-stage model to evaluate both trial and patient level surrogacy of correlated time-to-event endpoints using patient level data when multiple clinical trials are available. However, their maximum likelihood approach often suffers from numerical problems when different baseline hazards among trials and imperfect estimation of treatment effects are assumed. In Chapter 2, we propose performing the second-stage, trial-level evaluation of potential surrogates within a Bayesian framework, where we may naturally borrow information across trials while maintaining these realistic assumptions. Posterior distributions on surrogacy measures of interest may then be used to compare measures or make decisions regarding the candidacy of a specific endpoint. We perform a simulation study to investigate differences in estimation performance between traditional maximum likelihood and new Bayesian representations of common meta-analytic surrogacy measures, while assessing sensitivity to data characteristics such as number of trials, trial size, and amount of censoring. Furthermore, we present both frequentist and Bayesian trial-level surrogacy evaluations of time-to-recurrence for overall survival in two meta-analyses of adjuvant therapy trials in colon cancer. Based on these results, we recommend hierarchical Bayesian methods as an attractive alternative in the multi-trial evaluation of potential surrogate endpoints.

1.2 Bayesian Adaptive Trial for a Newly Validated Surrogate

The evaluation and validation of surrogate endpoints as primary endpoints in future clinical trials is an increasingly important research area, due to demands for

more efficient trials coupled with recent regulatory acceptance of some surrogates as 'valid.' However, little consideration has been given to how a trial which utilizes a newly-validated surrogate endpoint as its primary endpoint might be appropriately designed. In Chapter 3, we propose a novel Bayesian adaptive trial design that allows the new surrogate endpoint to play a dominant role in assessing the effect of an intervention, while remaining realistically cautious about its use. By incorporating multi-trial historical information on the validated relationship between the surrogate and clinical endpoints, then subsequently evaluating accumulating data against this relationship as the new trial progresses, we adaptively guard against an erroneous assessment of treatment based upon a truly invalid surrogate. So long as joint outcomes in the new trial seem plausible given similar historical trials, we proceed with trusting the surrogate endpoint as the primary endpoint, and do so adaptively–perhaps stopping the trial for early success or inferiority of the experimental treatment, or for futility. Otherwise, we discard the surrogate and switch adaptive determinations to the original primary endpoint. We use simulation to test the operating characteristics of this new design compared to a standard O'Brien-Fleming approach (O'Brien and Fleming, 1979), as well as the ability of our design to discriminate trustworthy from untrustworthy surrogates in hypothetical future trials. Furthermore, we investigate benefits using patient-level data from 18 adjuvant therapy trials in colon cancer, where disease-free survival is considered a newly-validated surrogate endpoint for overall survival.

1.3 Flowgraph Modeling for Surrogacy Evaluation

Flowgraphs are useful for modeling diseases that are well-described by multi-state models, but for which Markov assumptions do not hold and returns to previous disease states are possible. Furthermore, covariates such as treatment assignment may influence the distribution of waiting times between any two given states, and

to a differing degree for separate state transitions within the same system. In Chapter 4, we explore the construction and advantages of flowgraph models when used to describe progression through cancer disease states within clinical trials, where complicating factors such as returns to previous states and prolonged time periods in single states may be possible for some patients. Our goal, aided by a Bayesian approach, is improved parametric modeling of disease state transitions and compositions of transitions, resulting in more detailed and flexible estimation of treatment effects, hazard functions, and predictive densities of patient outcomes. With more sensitive parametric modeling of these quantities now possible, more realistic evaluations of potential surrogate endpoints such as disease-free survival for overall survival may be conducted. We describe possible flowgraph models for two recent clinical trials in colorectal cancer, and discuss how this modeling approach may be used to better understand potential surrogacy relationships.

CHAPTER TWO

Bayesian Adjusted R^2 for the Meta-Analytic Evaluation of Surrogate Time-to-Event Endpoints in Clinical Trials

2.1 Background

Surrogate endpoints are often desired in clinical trials when the primary clinical endpoint is costly to obtain, difficult to measure, or requires a long period of follow-up to observe. Controversy has surrounded the evaluation and use of potential surrogates since Prentice (1989) proposed a formal definition of surrogate endpoints and operational criteria for their validation. Despite ongoing and legitimate concerns of statisticians and clinicians alike, evaluation of potentially beneficial treatments in some settings remains impossible or implausible without the use of secondary endpoints that may occur sooner or more often among patients in a trial.

Fleming (1994) was one of the first to suggest that a *collection* of similar trials may be necessary to validate a surrogate endpoint. Hughes et al. (1995) echoed the sentiment, promoting meta-analyses as ideal for evaluating surrogacy across trials of unequal size and varied treatment effects. Early meta-analytic approaches involved modeling the association between treatment effects on the surrogate and true endpoints across trials, and predicting the effect of treatment on the clinical endpoint in a new trial given the treatment's effect on the surrogate and past information from similar trials. While many have noted the benefits of multi-trial or meta-analytic evaluation of potential surrogates, namely increased sample size and generalizability to future trials within a class of interventions, few have made such evaluations within a Bayesian framework.

The first paper involving a Bayesian meta-analysis for surrogate endpoint evaluation is that of Daniels and Hughes (1997), who considered a mixed effects model

for the trial-level association of treatment effects on the true clinical outcome and potential surrogate marker. However, their approach used only summary data, and thus did not explicitly model patient-level association between the two endpoints. Buyse and Molenberghs (1998) proposed validation of a surrogate endpoint at both patient and trial levels, for which they introduced the relative effect (RE) and adjusted association (AA) measures for the case of two binary endpoints. Later, Buyse et al. (2000) presented a new two-stage approach for modeling both individual and trial level surrogacy in the case of normal endpoints. The resulting quantities, R^2_{indiv} and R^2_{trial}, lie on the unit interval and represent the predictive abilities of the surrogate endpoint for the true endpoint, and the treatment effect on the surrogate for the treatment effect on the true endpoint, respectively. Burzykowski et al. (2001) then extended this approach to the case of time-to-event endpoints, where they proposed an adjusted trial-level surrogacy measure, R^2_{adj}, which takes estimation error of the treatment effects at the first stage into account at the second stage. However, in each of the case studies they considered, numerical problems cause R^2_{adj} to be unavailable unless common baseline hazards across trials are assumed.

In this paper, we propose bringing the advantages of the two-stage setting into a Bayesian framework at the second stage, where we may naturally borrow strength across trials in the evaluation of potential surrogate time-to-event endpoints while maintaining realistic assumptions. At the first stage, we model the truly bivariate nature of the candidate and surrogate endpoints at the individual level using copula models with trial-specific marginal distributions. At the second stage, we capture the association of the treatment effects on the surrogate and true endpoints at the trial level using Bayesian mixed effects models, arriving at posterior distributions on both the naive (unadjusted) and more realistic (adjusted) trial-level surrogacy measures. Through this Bayesian approach, one may use these posterior distributions for decision-making regarding the candidacy of a specific endpoint, or to make useful

probabilistic statements. For example, we may provide the probability given the observed data that R^2_{trial} is greater than some threshold of interest, or the probability that R^2_{trial} for one potential surrogate exceeds R^2_{trial} for another.

To motivate consideration of the Bayesian framework for evaluating trial-level surrogacy, we perform a simulation study to assess the sensitivity of frequentist and Bayesian trial-level surrogacy measures to underlying data characteristics such as "true" trial-level surrogacy, amount of censoring, number of trials, trial size, patient-level correlation of the true and surrogate endpoints, and range of treatment effects across trials. Furthermore, we perform classical and Bayesian evaluations of the surrogacy of time-to-recurrence (TTR) for overall survival (OS) in two meta-analyses of adjuvant therapy trials in colon cancer (Sargent et al., 2005).

The remainder of the paper evolves as follows. In Section 2.2, we review the popular existing two-stage model for trial-level surrogacy estimation in the case of time-to-event endpoints introduced by Burzykowski et al. (2001), discuss issues encountered with maximum likelihood estimation, and describe a Bayesian approach which may reliably used to obtain estimates of both unadjusted and adjusted measures. Section 2.3 presents the results of our simulation study, and we compare the classical and Bayesian approaches to trial-level surrogacy for two meta-analyses of adjuvant trials in colon cancer in Section 2.4. Finally, Section 2.5 concludes with a suggestion for modification of adjusted R^2_{trial} from that proposed by Burzykowski et al. (2001).

2.2 Two-Stage Model for Trial-Level Surrogacy

2.2.1 Review of Classical Approach

We begin with a review of the two-stage surrogacy model for time-to-event endpoints proposed by Burzykowski et al. (2001), which we will later bring into a Bayesian framework. Adopting established notation, we denote by the pair (T_{ij}, S_{ij})

the times to the true clinical endpoint and the potential surrogate endpoint, respectively, for the jth patient in the ith trial for $j = 1, ..., n_i$ and $i = 1, ..., N$. We denote by Z_{ij} the corresponding treatment assignment, where $Z_{ij} = 1$ for experimental and $Z_{ij} = 0$ for control. Note that one or both events of interest may be unobserved (right-censored) for some patients; if so, notation may be easily extended.

If S is to be considered as a surrogate for T, it is reasonable to assume that these endpoints may have nonzero correlation. Thus, a bivariate model seems appropriate to capture the patient-level association between the surrogate and true endpoints after adjusting for treatment. While many such bivariate survival models exist, a copula model (Genest and MacKay, 1986) is chosen for both its flexibility and the unique property that the margins do not depend on the specific choice of copula function. The joint survival function of S and T is then given by

$$F(s,t) = P(S_{ij} \geq s, T_{ij} \geq t) = C_\delta\{F_{S_{ij}}(s), F_{T_{ij}}(t)\}, \qquad s,t \geq 0, \tag{2.1}$$

where $(F_{S_{ij}}, F_{T_{ij}})$ are marginal survival functions and C_δ is the copula, a distribution function on $[0,1]^2$ having association parameter δ that describes the association between S_{ij} and T_{ij} at the patient level common to each trial.

Following Burzykowski et al. (2001), we assume proportional hazards in modeling the effect of treatment on the marginal distributions of S_{ij} and T_{ij} within the ith trial:

$$F_{S_{ij}}(s) = \exp\left\{-\int_0^s \lambda_{S_i}(x)\exp(\alpha_i Z_{ij})dx\right\} \text{ and } F_{T_{ij}}(t) = \exp\left\{-\int_0^t \lambda_{T_i}(x)\exp(\beta_i Z_{ij})dx\right\}. \tag{2.2}$$

Here, λ_{S_i} and λ_{T_i} are baseline hazard functions, and α_i and β_i are the effects of treatment Z on the surrogate and true endpoints, respectively, for trial i. A simplified version of this model, which assumes common baseline hazards across trials for each endpoint, is additionally considered in Burzykowski et al. (2001).

We choose to specify the marginal hazard functions for each trial parametrically using the Weibull distribution, thus matching the patient-level model considered by Burzykowski et al. (2001), but one may also leave the marginal hazard functions unspecified as in the Cox model (Cox, 1972). See Shih and Louis (1995) for an insightful discussion of the use of copulas with both parametric and semiparametric marginal survival models. While many choices of copula function are available to model the association between the two endpoints at the patient level, we focus on the copula model proposed by Clayton (1978) and used by Burzykowski et al. (2001), for which (2.1) becomes a proportional frailty model. The Clayton copula is given by

$$C_\delta(u,v) = (u^{1-\delta} + v^{1-\delta} - 1)^{1/(1-\delta)}, \qquad \delta > 1,$$

and generated by a Laplace transform $\phi_\delta(x) = (1+x)^{1/(1-\delta)}$ of the gamma distribution. With marginal Weibull models assumed for the effect of treatment on each endpoint, the joint survivor function under the Clayton copula, $F(s,t)$, becomes

$$[\exp\{-(1-\delta)(\gamma_{S_i}s)^{r_{S_i}}\exp(\alpha_i Z_{ij})\} + \exp\{-(1-\delta)(\gamma_{T_i}t)^{r_{T_i}}\exp(\beta_i Z_{ij})\} - 1]^{1/(1-\delta)},$$

where γ_{S_i} and r_{S_i} denote the trial-specific Weibull scale and shape parameters from the marginal model for the surrogate endpoint, and γ_{T_i} and r_{T_i} denote the corresponding Weibull parameters for the true endpoint. In this setting, S_{ij} and T_{ij} are positively associated when $\delta > 1$ and are independent when $\delta \to 1$.

At the second stage, we consider a random effects model for the trial-specific treatment effects (α_i, β_i) given by

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} a_i \\ b_i \end{bmatrix}, \tag{2.3}$$

where the random effects on the right-hand side of (2.3) are assumed to be normally

distributed with mean zero and covariance matrix

$$D = \begin{bmatrix} d_{aa} & d_{ab} \\ d_{ab} & d_{bb} \end{bmatrix}. \tag{2.4}$$

Under this model, which assumes perfect estimation of the trial-specific treatment effects when $(\widehat{\alpha}_i, \widehat{\beta}_i)$ are substituted for (α_i, β_i), a measure of trial-level surrogacy unadjusted for estimation error is given by $R^2_{un} = d^2_{ab}/d_{aa}d_{bb}$. Incidentally, R^2_{un} can be shown to be equivalent to the square of the correlation coefficient between treatment effects (α_i, β_i) across trials. However, the square of the sample correlation coefficient is well-known to be a biased estimator of the coefficient of determination. More importantly, the estimates $\widehat{\alpha}_i$ and $\widehat{\beta}_i$ computed from the first stage model are almost certainly not equal to the true treatment effects.

To overcome these bias issues, an adjusted trial-level modeling scheme based on developments by van Houwelingen et al. (2002) was proposed by Burzykowski et al. (2001). This model assumes the estimated treatment effects $\widehat{\alpha}_i$ and $\widehat{\beta}_i$ may be imperfectly measured at the first stage, and now follow the second stage model given by

$$\begin{bmatrix} \widehat{\alpha}_i \\ \widehat{\beta}_i \end{bmatrix} = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} + \begin{bmatrix} \epsilon_{ai} \\ \epsilon_{bi} \end{bmatrix}. \tag{2.5}$$

Here, the trial-specific estimation errors ϵ_{ai} and ϵ_{bi} are assumed to be jointly normally distributed with mean zero and trial-specific covariance

$$\Omega_i = \begin{bmatrix} \sigma_{aa,i} & \sigma_{ab,i} \\ \sigma_{ab,i} & \sigma_{bb,i} \end{bmatrix}. \tag{2.6}$$

The true trial-specific treatment effects (α_i, β_i) in (2.5) are still assumed to follow model (2.3) with covariance (2.4) for random effects (a_i, b_i). For the model parameters to be estimable, we follow Burzykowski et al. (2001) in assuming the covariance matrices Ω_i are known and equal to their estimates obtained from the

first stage copula model. An adjusted estimate of trial-level surrogacy is given by $R^2_{adj} = d^2_{ab}/d_{aa}d_{bb}$, similar to the unadjusted measure.

2.2.2 Issues with Maximum Likelihood Estimation

The first and second stage models above may be fitted in sequence for a given multi-trial dataset with time-to-event endpoints. In their advanced ovarian and advanced colorectal case studies, Burzykowski et al. (2001) first obtain maximum likelihood estimates for the copula models using a Newton-Raphson procedure, and then perform trial-level surrogacy estimation using standard software for mixed linear models. However, when we implemented their procedure for a number of simulated datasets, we found that the estimation algorithm for the first stage model occasionally fails to converge. Even when first stage estimates are available, it is very often the case that estimates of R^2_{adj} are unavailable in the second stage due to further numerical problems. Indeed, in both case studies presented by Burzykowski et al. (2001), estimates of R^2_{adj} were unavailable when baseline hazards were allowed to vary across trials, which is a critical element of an appropriate meta-analysis. In our experience, it seems to be the case that R^2_{adj} cannot be obtained in this setting when the number of trials or size of trials is too large. Specifically, among the scenarios we considered, we conservatively estimate that the procedure is very likely to fail for scenarios with more than 15 trials or more than 1500 patients in any trial, and remains considerably likely to fail otherwise.

With such issues in mind, Burzykowski and Abrahantes (2005) presented a robust alternative model for estimation of trial-level surrogacy in the presence of estimation error; however, unlike R^2_{adj}, the surrogacy estimate resulting from this model fails to have an analytical formula for the variance. Numerical problems abound for trial-level mixed and random effects models more complex than those described here, as is the case when a full random-effects model including trial-specific

intercepts is specified for the treatment effects. Options for model simplification through consideration of "dimensions" along which simplifying assumptions may be made, with the ultimate goal of obtaining model convergence and estimates of trial-level surrogacy, were discussed by Tibaldi et al. (2003). However, the issue remains that assumptions such as lack of error in estimation of the treatment effects and common baseline hazards across trials are unlikely to be reasonable in practice.

2.2.3 Bayesian Model for Second Stage

Rather than abandon or avoid reasonable modeling assumptions for fear of numerical problems with maximum likelihood estimation, we move the estimation error-adjusted surrogacy problem for survival endpoints into a Bayesian framework, with the goal of increasing stability of R^2-type estimates. We find that both unadjusted and adjusted Bayesian estimates are always available across a range of scenarios, including those where the desired classical adjusted measures often fail to exist.

Using the maximum likelihood estimates obtained by fitting the copula model at the first stage, we consider Bayesian estimation of the two trial-level models on the treatment effects described above: the naive unadjusted model given by (2.3) and (2.4) alone, and the adjusted mixed-effects model accounting for estimation error given by (2.3) - (2.6).

In fitting the naive second-stage model to the *estimated* treatment effects from the copula, we assume

$$\begin{bmatrix} \widehat{\alpha}_i \\ \widehat{\beta}_i \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, D = \begin{bmatrix} d_{aa} & d_{ab} \\ d_{ab} & d_{bb} \end{bmatrix} \right),$$

and consider a vague multivariate normal prior on the mean treatment effects (α, β)

with mean zero and precision

$$P = \begin{bmatrix} 1.0 \times 10^{-6} & 0 \\ 0 & 1.0 \times 10^{-6} \end{bmatrix}.$$

We place a similarly vague Wishart(ρ, M) prior on the normal precision matrix D^{-1}, with $M = P$ and $\rho = 2$. After Markov chain Monte Carlo (MCMC) estimation of posterior distributions for the model parameters, a posterior on R^2_{un} may be obtained from the formula $R^2_{un} = d^2_{ab}/d_{aa}d_{bb}$.

The second-stage model adjusted for estimation error of the treatment effects $\widehat{\alpha}_i$ and $\widehat{\beta}_i$ is given by

$$\begin{bmatrix} \widehat{\alpha}_i \\ \widehat{\beta}_i \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}, \Omega_i = \begin{bmatrix} \sigma_{aa,i} & \sigma_{ab,i} \\ \sigma_{ab,i} & \sigma_{bb,i} \end{bmatrix} \right) \tag{2.7}$$

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} a_i \\ b_i \end{bmatrix} \tag{2.8}$$

$$\begin{bmatrix} a_i \\ b_i \end{bmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, D = \begin{bmatrix} d_{aa} & d_{ab} \\ d_{ab} & d_{bb} \end{bmatrix} \right), \tag{2.9}$$

where the error precisions Ω_i^{-1} are fixed at their first-stage maximum likelihood estimates from the copula model, following Burzykowski et al. (2001). This model assumes $(\widehat{\alpha}_i, \widehat{\beta}_i)$ and $(\widehat{\alpha}_j, \widehat{\beta}_j)$ are independent for $i \neq j$, as it is not identifiable otherwise. However, this is not an unrealistic assumption, as the between-trial elements of the Hessian matrix estimated at the copula stage are equal to zero. We again consider a vague multivariate normal prior on the mean treatment effects (α, β) and place a vague Wishart prior on D^{-1}, the precision of the random effects. A posterior for the adjusted surrogacy quantity R^2_{adj} is still given from the posterior for D by the formula $R^2_{adj} = d^2_{ab}/d_{aa}d_{bb}$, where D denotes the covariance of the random effects in the adjusted model.

It should be noted that the model for R^2_{adj} developed by Burzykowski et al. (2001) and presented as (2.3) - (2.6) cannot be implemented using standard Bayesian computing packages without slight modification. As such, we introduce the variability of the copula-estimated treatment effects $(\widehat{\alpha}_i, \widehat{\beta}_i)$ due to the random-effects covariance D through the model for the "true" treatment effects (α_i, β_i), as described in (2.7) - (2.9) above.

2.3 Simulation Study

All simulations in this paper were performed with 500 iterations. At each iteration, we use a two-stage data generation procedure to produce correlated Weibull endpoints (S_{ij}, T_{ij}) for each of the n_i patients within trial i, $i = 1, ..., N$. Next, using the generated data, we fit the patient-level copula models to obtain maximum likelihood estimates of the trial-specific treatment effects and corresponding precision matrices. These estimates are used to fit the second-stage models and arrive at frequentist and Bayesian estimates of trial-level surrogacy. In particular, for each scenario, we compute posterior distributions for R^2_{un} and R^2_{adj}, as well as maximum likelihood estimates and standard errors of R^2_{un} for comparison. Frequentist estimates of R^2_{adj} are excluded from this simulation study, as they are almost always numerically unavailable.

2.3.1 Data Generation

At the first stage of data generation, we obtain the regression coefficient vector $(\mu_{S_i}, \mu_{T_i}, \alpha_i, \beta_i)$ for trial i from a multivariate normal distribution:

$$\begin{bmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \end{bmatrix} \sim MVN \left(\begin{bmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{bmatrix}, \Sigma_{trial} = \begin{bmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{bmatrix} \right). \qquad (2.10)$$

In determining the mean and covariance parameter elements of (2.10), we choose a value of the true trial-level R^2 from $0 < R^2_{trial} < 1$, which in turn determines a fixed value for d_{ab} in the covariance Σ_{trial} by the relationship $d_{ab} = \sqrt{R^2_{trial} d_{aa} d_{bb}}$. This is due to the relationship $R^2_{trial} = d^2_{ab}/d_{aa}d_{bb}$ used for surrogacy estimation. Additional components of the multivariate normal mean and covariance above are based on estimates from ACCENT, a collection of adjuvant therapy trials for colon cancer (Sargent et al., 2005, 2007), for which we consider TTR as a potential surrogate for OS.

At the second stage, we generate correlated Weibull survival times for each of the n_i patients in trial i. First, we use the fact that the Weibull distribution can be expressed as a scale mixture of half-normals (Cowles, 2004). For example, if λ and Y^* are independent random variables such that $\lambda \sim Exp(1)$, $0 < \lambda < \infty$, and $Y^* \sim$Truncated$N(0,1)$, $0 < Y^* < \infty$, then $T = \delta(Y^* \times \sqrt{2\lambda})^{1/\gamma}$ with $0 < \delta < \infty$, $0 < \gamma < \infty$, will follow a Weibull distribution:

$$f_T(t) = \frac{\gamma}{\delta}\left(\frac{t}{\delta}\right)^{\gamma-1} \exp\left(-\left(\frac{t}{\delta}\right)^{\gamma}\right), 0 < t < \infty.$$

In the expression above, δ contains the regression component and γ is a scale parameter. In our case, the regression components are given by $\delta_{S_{ij}} = \exp(\mu_{S_i} + \alpha_i z_{ij})$ and $\delta_{T_{ij}} = \exp(\mu_{T_i} + \beta_i z_{ij})$, where z_{ij} is the treatment indicator for the jth individual in the ith trial.

A factor of investigation in our simulation study is the effect of strongly correlated overall survival (OS) and time-to-recurrence (TTR) versus uncorrelated OS and TTR on trial-level surrogacy. To construct death times T_{ij} and recurrence times S_{ij} which are strongly correlated at the patient level, we first generate the random variates $Y^*_{ij} \sim$ Truncated $N(0,1)$, $\lambda_{S_{ij}} \sim Exp(1)$, and $\lambda_{T_{ij}} \sim Exp(1)$ for the ith patient in the jth trial. From these, patient-specific event times are given by $S_{ij} = \delta_{S_{ij}}(Y^*_{ij} \times \sqrt{2\lambda_{S_{ij}}})^{1/\gamma_S}$ and $T_{ij} = \delta_{T_{ij}}(Y^*_{ij} \times \sqrt{2\lambda_{T_{ij}}})^{1/\gamma_T}$. In this case, use of a common value of Y^*_{ij} induces a positive correlation between S_{ij} and

T_{ij}. Uncorrelated event times are constructed by first drawing separate random variates $Y^*_{S_{ij}} \sim Truncated\, N(0,1)$ and $Y^*_{T_{ij}} \sim$ Truncated $N(0,1)$, which are then used with the variates $\lambda_{S_{ij}} \sim Exp(1)$ and $\lambda_{T_{ij}} \sim Exp(1)$ to produce event times $S_{ij} = \delta_{S_{ij}} \left(Y^*_{S_{ij}} \times \sqrt{2\lambda_{S_{ij}}} \right)^{1/\gamma_S}$ and $T_{ij} = \delta_{T_{ij}} \left(Y^*_{T_{ij}} \times \sqrt{2\lambda_{T_{ij}}} \right)^{1/\gamma_T}$. Endpoints generated to be uncorrelated in our scenarios in fact have a correlation very near zero, while endpoints constructed to be strongly correlated have a correlation near 0.70.

For this paper, we generated data sets according to the scenarios of interest found in Table 2.1. Due to computation time and numerous potential combinations of factors, we begin our investigation by assessing each of the factors separately, while holding all other factors fixed at "optimal" levels given by the first element in each row above: $R^2_{trial} = 0.90$, 30 trials, 2000 patients per trial, no censoring, strong correlation, and a large range of hazard ratios. Throughout, we assume two treatment groups with equal allocation. For scenarios involving censoring, we assume non-informative, uniform censoring, and we control the amount of censoring on OS *within each trial* through the upper bound, $kT_{0.5}$ of the $Uniform(0, kT_{0.5})$ distribution, where $T_{0.5}$ is the median observed OS time in the trial and k is a chosen constant. Because we control the ratio of median TTR times to median OS times in the data generation process, the censoring rate for TTR is then naturally derived from the censoring rate for OS. We control the range of hazard ratios across trials by changing d_{aa} and d_{bb} by a multiplicative factor at the first stage of data generation given by (2.10). This produces the scenarios given in the last row of Table 2.1, where we are interested to compare surrogacy across the cases when our trials yield a large range of hazard ratios including 1, a small range of hazard ratios including 1, and a small range of hazard ratios excluding 1.

Table 2.1: Factors of interest and corresponding levels considered in the simulation study.

Factors	Levels Explored in Simulations
Trial-level R^2 for TTR, OS	0.90, 0.60, 0.20
Number of trials	30, 15, 5
Sample sizes within trials (equal)	2000, 1000, 500
Unequal trial sizes	100% at $n = 2000$
	33.3% each at $n = (500, 1000, 2000)$
	50% each at $n = (500, 2000)$
Censoring rate for OS	0%, 30%, 70%
Individual-level correlation	Strong (correlation near 0.70)
	Weak (correlation near 0)
Range of hazard ratios across trials	Large (includes $HR = 1$)
	Small (includes $HR = 1$)
	Small (excludes $HR = 1$)

2.3.2 Estimation of Two-Stage Model

We estimate the first-stage model parameters by first constructing the likelihood function, assuming Clayton copulas with endpoint-specific Weibull marginal models for the true and surrogate endpoints within each trial, and a common copula association parameter across trials. The likelihood is also rewritten for programming purposes in terms of appropriately transformed parameters, so that each parameter to be estimated has support over the entire real line. Maximum likelihood estimates and numerically approximated information matrices may then be obtained for use in second-stage frequentist and Bayesian estimation of trial-level surrogacy.

As we would like to compare maximum likelihood estimates of trial-level surrogacy with posterior estimates using the Bayesian approach, we obtain second-stage frequentist estimates of R^2_{un} by considering a generalized least squares representation of the unadjusted model. For Bayesian estimation, we use the *estimated* treatment effects and covariances across trials saved from the copula stage of estimation to fit the unadjusted and adjusted second-stage models. In our MCMC algorithm, initial values for mean parameters were set to zero, and identity matrices were used as initial values for covariance parameters. After checking that convergence diagnostics

such as negligible autocorrelation were satisfied for each scenario, we saved posterior sample chains of length 10,000 with a burn-in of 1,000 iterations. Because posterior distributions for R^2 are often skewed, we consider both the posterior mean and the posterior median as estimators of trial-level surrogacy.

2.3.3 Results

Simulation results for the primary factors of interest listed in Table 2.1 are presented numerically in Tables 2.2 and 2.3 and graphically in Figures 2.1 and 2.2. Maximum likelihood estimates of R^2_{un}, denoted by $\widehat{R}^2_{un}$ in the tables, were used with their estimated standard errors (Hotelling, 1953) across iterations to obtain estimates of bias, mean-squared error (MSE), and coverage of 95% confidence intervals. For the Bayesian unadjusted posterior means and medians, denoted by $E(R^2_{un}|D)$ and $Md(R^2_{un}|D)$, respectively, we present estimates of bias, MSE, and coverage of equal-tailed posterior 95% credible intervals for comparison with their frequentist counterparts. We present similar summaries of estimation performance for the posterior mean and median, $E(R^2_{adj}|D)$ and $Md(R^2_{adj}|D)$, of the trial-level surrogacy measure based on the Bayesian model adjusting for estimation error. Box plots of $\widehat{R}^2_{un}$, $E(R^2_{un}|D)$, $Md(R^2_{un}|D)$, $E(R^2_{adj}|D)$ and $Md(R^2_{adj}|D)$ are presented for assessment of bias and variability across the 500 datasets generated at each simulation setting.

Based on the results for the individual factors, many of which demonstrate an advantage for the Bayesian adjusted measure as a factor is worsened (i.e., smaller trials or smaller range of treatment effects), we performed a group of "secondary simulations" where many factors were worsened simultaneously. In particular, we consider simulated meta-analyses with 500 patients per trial, 70% censoring for OS, weak patient-level correlation, and a small range of hazard ratios including $HR = 1$. The only factor we varied was number of trials, again considering scenarios with 30,

15, and 5 trials. These results are presented as the last section of Table 2.3 and in the lower right plot of Figure 2.2.

In the simulation results presented in Tables 2.2 and 2.3, it is clear that Bayesian R^2_{adj} demonstrates the best estimation performance across all the scenarios considered. The greatest advantages in terms of bias and MSE for this measure are exhibited when there are few trials, small trials, or trials of varying size, when the censoring rate for OS is high, when patient-level correlation between OS and TTR is low, or when the range of treatment effects across trials is small. Coverage of 95% equal-tailed intervals is also best for adjusted R^2 in these scenarios, with the exception of $N = 5$ trials. Upon further investigation, we find that the precision of adjusted R^2 estimates may become too small relative to their bias for $N = 5$, producing intervals that exclude the true value. The posterior mean and median of the adjusted surrogacy measures also perform better than the unadjusted measures across all levels of true trial-level surrogacy, although their benefits do not change as a function of surrogacy strength. The Bayesian unadjusted measures, based on models with relatively flat priors, perform similarly to the maximum likelihood estimates. Small gains for the Bayesian approach to unadjusted surrogacy estimation may be noted in most cases, however, when comparing $\widehat{R}^2_{un}$ to $Md(R^2_{un}|D)$. Indeed, within the Bayesian approach, the posterior median seems to be a better estimator of trial-level surrogacy (unadjusted or adjusted) than the posterior mean when estimates are biased low, due most likely to the fact that posteriors for R^2 are often skewed. The graphical results of primary simulations in Figures 2.1 and 2.2 provide additional motivating support for Bayesian R^2_{adj}, easily showing where its gains are largest relative to the unadjusted measures.

In Figure 2.1, noticeable improvements in bias are demonstrated by Bayesian R^2_{adj} measures over the unadjusted measures for a high level of censoring (70%) and small trial size (500 patients per trial), with more modest improvements for low

Table 2.2: Impact of trial characteristics on maximum likelihood estimates, $\widehat{R}^2$, Bayesian posterior means, $E(R^2|D)$, and Bayesian posterior medians, $Md(R^2|D)$, for unadjusted and adjusted measures of surrogacy of TTR for OS.

Surrogacy Estimate	Bias	Coverage	MSE	Bias	Coverage	MSE	Bias	Coverage	MSE
				True Trial-Level Surrogacy					
		R^2 = 0.90			*R^2 = 0.60*			*R^2 = 0.20*	
$\widehat{R}^2_{un}$	-0.029	0.904	0.003	-0.015	0.946	0.015	0.026	0.952	0.016
$E(R^2_{un}\|D)$	-0.032	0.908	0.004	-0.015	0.934	0.014	0.042	0.940	0.015
$Md(R^2_{un}\|D)$	-0.026	0.908	0.003	-0.008	0.934	0.014	0.031	0.940	0.017
$E(R^2_{adj}\|D)$	-0.024	0.912	0.003	-0.014	0.930	0.014	0.038	0.940	0.015
$Md(R^2_{adj}\|D)$	-0.017	0.912	0.003	-0.005	0.930	0.015	0.028	0.940	0.017
				Number of Trials					
		N = 30			*N= 15*			*N = 5*	
$\widehat{R}^2_{un}$	-0.029	0.904	0.003	-0.032	0.948	0.007	-0.049	0.974	0.029
$E(R^2_{un}\|D)$	-0.032	0.908	0.004	-0.037	0.934	0.007	-0.057	0.844	0.026
$Md(R^2_{un}\|D)$	-0.026	0.908	0.003	-0.024	0.934	0.006	-0.030	0.844	0.024
$E(R^2_{adj}\|D)$	-0.024	0.912	0.003	-0.027	0.934	0.006	0.011	0.552	0.017
$Md(R^2_{adj}\|D)$	-0.017	0.912	0.003	-0.014	0.934	0.005	0.031	0.552	0.015
				Trial Size					
		2000			*1000*			*500*	
$\widehat{R}^2_{un}$	-0.029	0.904	0.003	-0.039	0.884	0.004	-0.074	0.698	0.009
$E(R^2_{un}\|D)$	-0.032	0.908	0.004	-0.041	0.880	0.004	-0.077	0.708	0.010
$Md(R^2_{un}\|D)$	-0.026	0.908	0.003	-0.035	0.880	0.004	-0.069	0.708	0.009
$E(R^2_{adj}\|D)$	-0.024	0.912	0.003	-0.025	0.928	0.003	-0.046	0.874	0.006
$Md(R^2_{adj}\|D)$	-0.017	0.912	0.003	-0.018	0.928	0.003	-0.037	0.874	0.005
				Censoring Rate					
		0%			*30%*			*70%*	
$\widehat{R}^2_{un}$	-0.029	0.904	0.003	-0.056	0.802	0.007	-0.093	0.598	0.013
$E(R^2_{un}\|D)$	-0.032	0.908	0.004	-0.059	0.806	0.007	-0.096	0.616	0.013
$Md(R^2_{un}\|D)$	-0.026	0.908	0.003	-0.052	0.806	0.006	-0.088	0.616	0.012
$E(R^2_{adj}\|D)$	-0.024	0.912	0.003	-0.045	0.870	0.006	-0.062	0.834	0.008
$Md(R^2_{adj}\|D)$	-0.017	0.912	0.003	-0.037	0.870	0.005	-0.053	0.834	0.007
				Mix of Trial Sizes					
		2000			*(500, 1000, 2000)*			*(500, 2000)*	
$\widehat{R}^2_{un}$	-0.029	0.904	0.003	-0.045	0.872	0.005	-0.047	0.848	0.005
$E(R^2_{un}\|D)$	-0.032	0.908	0.004	-0.048	0.880	0.005	-0.050	0.850	0.006
$Md(R^2_{un}\|D)$	-0.026	0.908	0.003	-0.042	0.880	0.004	-0.043	0.850	0.005
$E(R^2_{adj}\|D)$	-0.024	0.912	0.003	-0.028	0.924	0.003	-0.028	0.912	0.004
$Md(R^2_{adj}\|D)$	-0.017	0.912	0.003	-0.020	0.924	0.003	-0.020	0.912	0.003

Table 2.3: Impact of trial characteristics on maximum likelihood estimates, $\widehat{R}^2$, Bayesian posterior means, $E(R^2|D)$, and Bayesian posterior medians, $Md(R^2|D)$, for unadjusted and adjusted measures of surrogacy of TTR for OS.

Surrogacy Estimate	Bias	Coverage	MSE	Bias	Coverage	MSE	Bias	Coverage	MSE
Range of Hazard Ratios for OS Across Trials									
	Large, Includes $HR = 1$			*Small, Includes* $HR = 1$			*Small, Excludes* $HR = 1$		
$\widehat{R}^2_{un}$	-0.029	0.904	0.003	-0.208	0.118	0.053	-0.214	0.116	0.056
$E(R^2_{un}\|D)$	-0.032	0.908	0.004	-0.210	0.120	0.053	-0.217	0.118	0.057
$Md(R^2_{un}\|D)$	-0.026	0.908	0.003	-0.202	0.120	0.050	-0.208	0.118	0.053
$E(R^2_{adj}\|D)$	-0.024	0.912	0.003	-0.110	0.574	0.030	-0.125	0.562	0.034
$Md(R^2_{adj}\|D)$	-0.017	0.912	0.003	-0.093	0.574	0.028	-0.109	0.562	0.031
Individual-Level Correlation									
		Strong			*Weak*				
$\widehat{R}^2_{un}$	-0.029	0.904	0.003	-0.051	0.836	0.005			
$E(R^2_{un}\|D)$	-0.032	0.908	0.004	-0.054	0.852	0.006			
$Md(R^2_{un}\|D)$	-0.026	0.908	0.003	-0.047	0.852	0.005			
$E(R^2_{adj}\|D)$	-0.024	0.912	0.003	-0.024	0.922	0.003			
$Md(R^2_{adj}\|D)$	-0.017	0.912	0.003	-0.016	0.922	0.003			
Secondary Simulations - Number of Trials									
		N=30			*N=15*			*N=5*	
$\widehat{R}^2_{un}$	-0.843	0	0.715	-0.810	0	0.666	-0.646	0.490	0.483
$E(R^2_{un}\|D)$	-0.815	0	0.668	-0.760	0	0.585	-0.554	0.222	0.346
$Md(R^2_{un}\|D)$	-0.835	0	0.702	-0.792	0	0.637	-0.587	0.222	0.402
$E(R^2_{adj}\|D)$	0.026	0.358	0.020	-0.105	0.690	0.045	-0.325	0.964	0.120
$Md(R^2_{adj}\|D)$	0.065	0.358	0.019	-0.028	0.690	0.030	-0.283	0.964	0.104

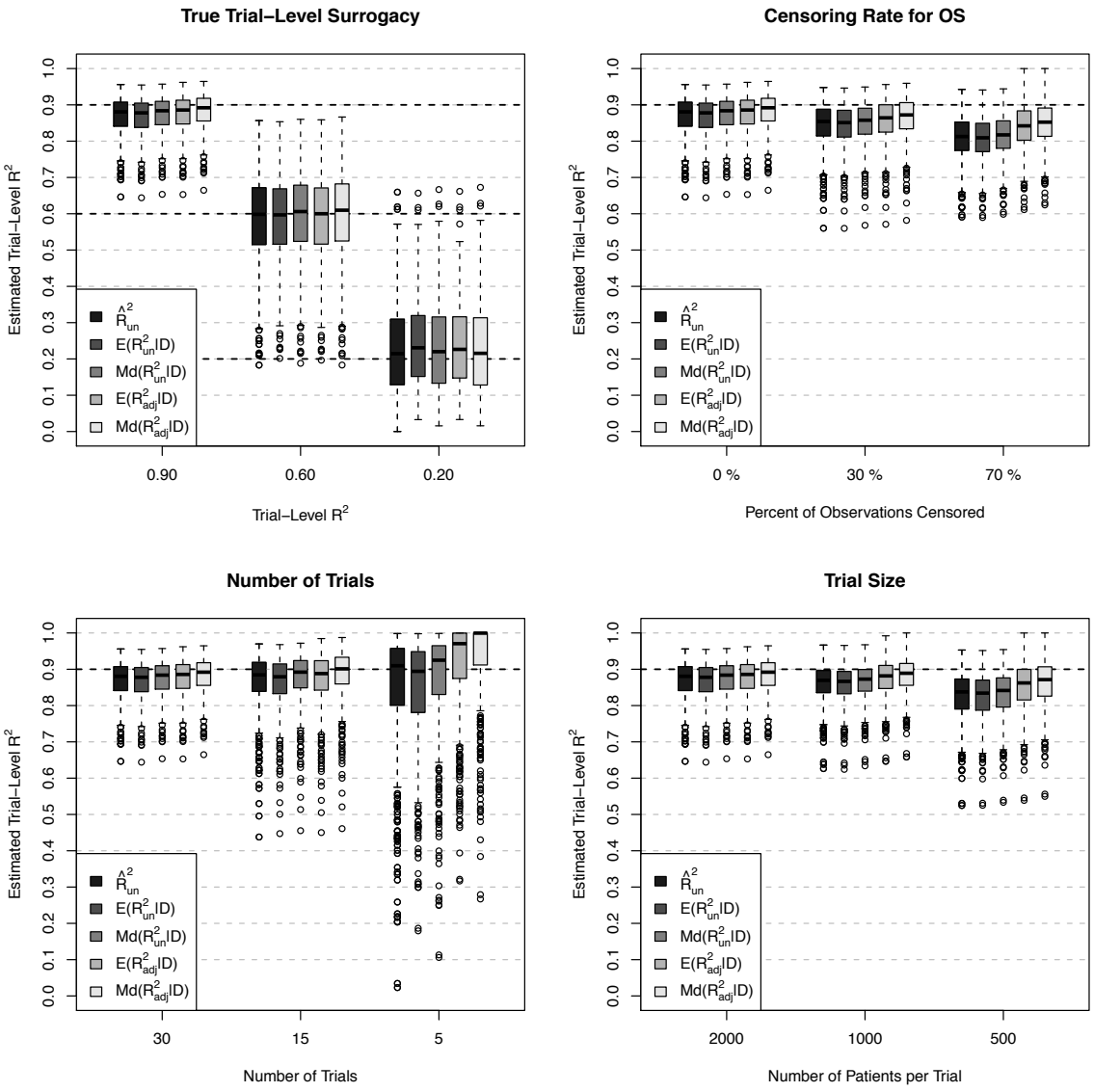

Figure 2.1: Boxplots comparing estimation performance of trial-level surrogacy measures for primary simulation settings.

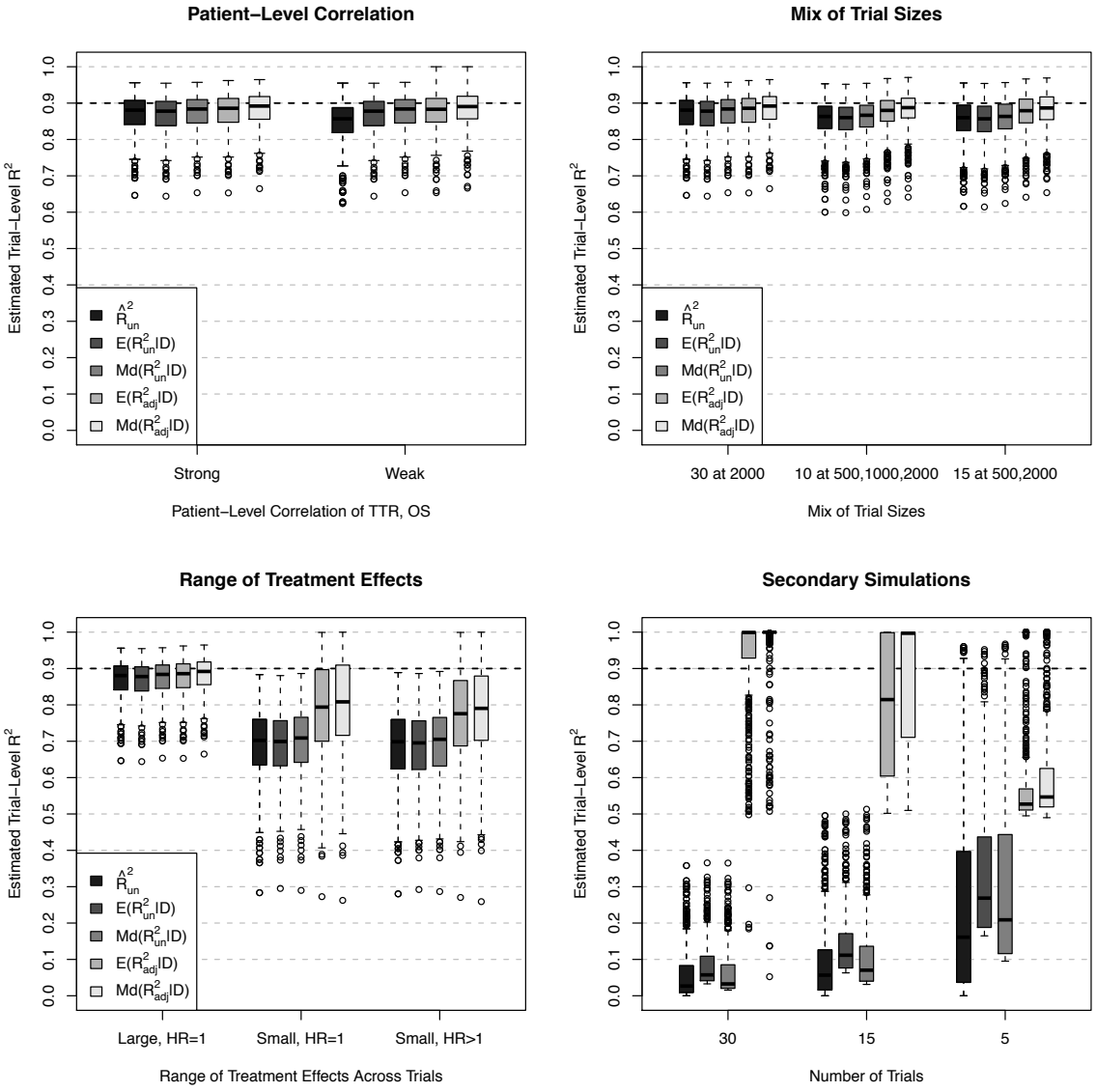

Figure 2.2: Boxplots comparing estimation performance of trial-level surrogacy measures for primary and secondary simulation settings.

censoring and moderate trial size. In Figure 2.2, Bayesian R^2_{adj} shows less bias for disparate trial sizes (15 trials each at $n = (500, 2000)$), and is the irrefutable winner when a small range of treatment effects is present across trials–both when $HR = 1$ is included and when $HR = 1$ is excluded.

Moving away from the best-case scenarios toward those where multiple trial characteristics are worsened simultaneously, we find tremendous advantages for Bayesian R^2_{adj}. As expected, the increased bias and variability evident for the unadjusted measures when only one trial characteristic is weakened becomes magnified when factors are weakened in combination. The extremely high bias and MSE observed for $\widehat{R}^2_{un}$ is only slightly improved by consideration of the Bayesian counterparts $E(R^2_{un}|D)$ and $Md(R^2_{un}|D)$, and all three measures have coverage equal to zero for analyses based on 15 or 30 trials. Evidently, as shown in the lower-right plot of 2.2, R^2_{un} becomes useless to detect a truly high level of surrogacy in meta-analyses with far from ideal trial characteristics. This is an important discovery, as these scenarios are arguably the most realistic considered in our study.

The estimation performance of Bayesian R^2_{adj} for these combination scenarios, on the other hand, is only mildly affected by the combined worsening of all factors. Bias and MSE are drastically lower for both $E(R^2_{adj}|D)$ and $Md(R^2_{adj}|D)$ compared to the unadjusted measures, and while coverage is not acceptably close to 95% for 15 or 30 trials, it is far from zero. In the very worst scenario considered in this paper – only 5 small trials with high censoring – Bayesian R^2_{adj} suffers considerably in terms of bias and MSE, but not to the extent of the unadjusted measures. Furthermore, coverages for $E(R^2_{adj}|D)$ and $Md(R^2_{adj}|D)$ are near 95%. Box plots of the surrogacy estimates for the secondary simulations are displayed in the final window of Figure 2.2. While the adjusted measures are certainly located nearer to the truth of $R^2 = 0.90$, it is interesting to note how the estimates pile up near 1 for a large number of trials.

This anomaly will become especially relevant when we estimate surrogacy for the ACCENT trials in Section 2.4.

2.4 Application to Adjuvant Trials in Colon Cancer

2.4.1 ACCENT Data

In this section, we illustrate our methods with an example from colorectal cancer, the third most common cancer in the United States with approximately 145,000 new cases diagnosed every year (Sargent et al., 2007). When no interventions are administered to patients with node-positive disease after primary resection, approximately half will experience relapse and eventually die as a result of their disease. Sargent et al. (2005) demonstrated that disease-free survival (DFS) with a median of 3 years follow-up is a valid surrogate for overall survival (OS) with a median of 5 years follow-up in the adjuvant setting, based on a variety of graphical and statistical approaches. Their trial-level surrogacy quantities, obtained by fitting marginal (independent) Cox models to each endpoint within each trial, included the coefficient of determination from the weighted regression of hazard ratios for OS onto the hazard ratios of DFS across trials. Later, Sargent et al. (2007) evaluated the surrogacy of DFS at various lengths of median follow-up for 5 years of median follow-up on OS, considering stage II and stage III disease both separately and in pooled analyses. However, these previous meta-analyses for trial-level surrogacy did not adjust for error in estimation of the treatment effects, due to the fact that these results were numerically unavailable.

To address this issue, we demonstrate both unadjusted and adjusted surrogacy estimation on trials provided by the Adjuvant Colon Cancer End Points (ACCENT) Group (Sargent et al., 2005, 2007), which contain individual patient data from 18 randomized phase II and phase III trials for adjuvant therapy in colon cancer. These

trials were conducted from 1977 to 1999, and collectively include 20,898 patients assigned to 43 treatment arms, composed of 34 active treatment arms (with at least one fluorouracil (FU)-based chemotherapy arm per trial) and 9 surgery-only arms. In our analysis of these data, we consider only those ACCENT trials where surgery-only arms serve as control to be compared with active treatment (trials conducted from 1977 to 1989). We maintain the natural ordering of the control and treatment arms in the original trial designs, and estimate surrogacy based on the $N = 9$ two-arm comparisons or trial units which meet our criteria. Our analysis is quite different from previous analyses of these data by Sargent et al. (2005), where all available trials and pairwise comparisons to the control arm within each trial were used to evaluate surrogacy of DFS for OS based on $N = 25$ trial-level units, and where additional censoring was imposed to construct specific median follow-up times for each endpoint. In our analysis, we use all available patient follow-up for each endpoint without imposing censoring beyond what exists in the data, and furthermore, we are only interested in the surrogacy of TTR for OS in the 9 two-arm trials with surgery-only control arms.

For comparison, we also investigate whether the level of TTR surrogacy for OS observed in the older ACCENT trials is consistent with surrogacy of TTR for OS in a collection of newer ACCENT trials conducted from 1997 to 2002 (Sargent et al., 2011), where all patients were treated with chemotherapy after surgical resection. This meta-analysis is based on individual-level data from 12,676 stage II and III colon cancer patients enrolled on 6 two-arm phase III adjuvant trials with a total of 12 distinct treatment arms. While the control arms in the older ACCENT trials considered in this paper are based on treatment with surgery alone, the 6 control arms in the new ACCENT trials are based on standard-of-care 5-FU/LV regimens. Experimental arms for the newer trials include two arms each of 5-FU/LV in combination with oxaliplatin, 5-FU/LV in combination with irinotecan, and an oral

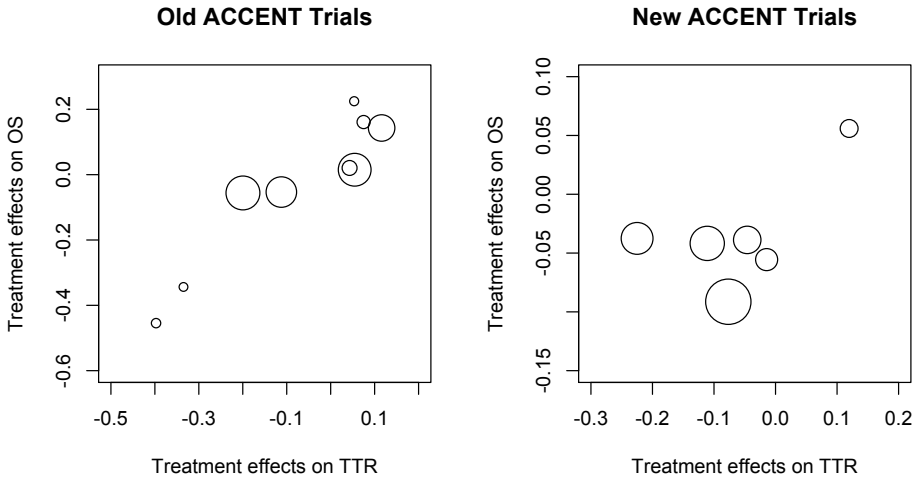

Figure 2.3: Estimated treatment effects on OS vs TTR for the old ACCENT trials and new ACCENT trials, weighted by trial size.

fluoropyrimidine. Furthermore, the median time from recurrence to death in the newer ACCENT trials is 2 years, in contrast to a median recurrence-to-death of 12 months among the older ACCENT trials.

2.4.2 Analysis of TTR Surrogacy for OS

We may begin these analyses with a visual assessment of the relationship between copula-estimated treatment effect estimates for OS and TTR across trials by scatterplots, presented in Figure 2.3 for both older ACCENT and newer ACCENT datasets. There is a strong positive linear relationship evident between the treatment effects on OS and the treatment effects on TTR for the 9 older ACCENT trials with surgery-only arms, suggesting that the effect of treatment on TTR has prognostic value for the effect of treatment on OS. While only 6 new ACCENT trials are available, the relationship between the treatment effects in this case seems decidedly less strong, and perhaps slightly nonlinear in nature. Nonetheless, we proceed with surrogacy estimation for each meta-analysis, assuming the treatment effects are linearly related across trials.

Just as for the simulation study presented in Section 2.3, we fit the first-stage Weibull copula regression models to each meta-analysis and use the resulting treatment effect estimates and precision matrices to fit both frequentist and Bayesian second-stage models. We obtain maximum likelihood estimates and standard errors of R^2_{un}, as well as corresponding 95% confidence intervals, while similar estimates for R^2_{adj} continue to be unavailable. As in the simulation study, we evaluate posterior quantities based on MCMC chains of length 10,000 plus 1,000 burn-in iterations. Bayesian posterior means and standard deviations for R^2_{un} and R^2_{adj} were then computed, along with concordance rates obtained by leave-one-out Bayesian prediction of the effect of treatment on OS in a new trial given the treatment's effect on TTR. In this sense, the rate of concordance is given by the percentage of trials for which the observed treatment effect on OS lies within the 95% equal-tailed credible interval for the treatment effect when it is assumed unknown. For simplicity of presentation, and due to posterior medians being uniformly higher than posterior means for both R^2_{un} and R^2_{adj} in each meta-analysis, we consider only posterior means as more "conservative" point estimates of true underlying surrogacy. Lastly, to demonstrate a more general advantage of the Bayesian approach, we compute the posterior probability that trial-level R^2 is greater than a chosen threshold of 0.90 for both unadjusted and adjusted measures within each ACCENT meta-analysis. These results are given in Table 2.4.

2.4.3 Results: Unadjusted R^2

When considering the older ACCENT trials with surgery-only control arms and all available patient follow-up, we find both Bayesian and classical estimates of R^2_{un} are slightly less than 0.90 with standard errors near 0.08, indicating that TTR would be a reasonably good surrogate endpoint for OS in trials of this kind. Both 95% confidence intervals and 95% credible sets for R^2_{un} exclude surrogacy levels less

Table 2.4: MLEs / posterior means, standard errors, 95% equal-tailed intervals, Bayesian concordance, and Bayesian posterior probabilities for the surrogacy of TTR for OS in the old and new ACCENT trials.

Estimator	Estimate	SE	95% Interval	Concord	$P(R^2 > 0.9)$
			9 Old ACCENT Trials		
$\widehat{R}^2_{un}$	0.882	0.078	(0.535, 0.975)	-	-
$E(R^2_{un}\|D)$	0.872	0.086	(0.647, 0.972)	1.00	0.458
$\widehat{R}^2_{adj}$	NA	NA	NA	-	-
$E(R^2_{adj}\|D)$	0.932	0.184	(0.187, 0.999)	0.56	0.867
			6 New ACCENT Trials		
$\widehat{R}^2_{un}$	0.377	0.342	(0.155, 0.905)	-	-
$E(R^2_{un}\|D)$	0.416	0.248	(0.006, 0.858)	0.67	0.010
$\widehat{R}^2_{adj}$	NA	NA	NA	-	-
$E(R^2_{adj}\|D)$	0.935	0.179	(0.213, 0.999)	1.00	0.873

than 0.50, suggesting that poor surrogacy of TTR for OS in such trials is unlikely. However, the posterior probability that R^2_{un} exceeds a "high surrogacy" threshold of 0.90 is only 0.458, given the observed data. The Bayesian predictive concordance of observed and predictive treatment effects is equal to 1.00 for the unadjusted trial-level model, suggesting that 95% posterior predictive intervals for the effect of treatment on OS are likely to contain the true treatment effect in a new trial, before this effect might be observed.

The newer ACCENT trials, however, yield much lower estimates of unadjusted trial-level surrogacy. Specifically, surrogacy of TTR for OS drops to 0.377 under maximum likelihood estimation and to 0.416 for the Bayesian model. These point estimates indicate poor surrogacy of TTR for OS in the new ACCENT trials, and we also note that they are accompanied by much larger standard errors than estimates from the older trials. Both 95% confidence intervals and 95% credible intervals for R^2_{un} in the newer trials are so wide as to suggest both very poor and good surrogacy are possible. However, the posterior probability that R^2_{un} exceeds 0.90 is very low at 0.010, meaning that very high levels of surrogacy are unlikely. In combination,

these results advise that TTR may be a poorer candidate surrogate for OS in newer adjuvant trials than in older adjuvant trials, and furthermore, R^2_{un} may be a poor surrogacy measure to use in evaluation of potential surrogates for new trials in colon cancer. Bayesian concordance of observed and predicted treatment effects on OS based on the unadjusted trial-level model is also lower for the new ACCENT data.

2.4.4 Results: Bayesian Adjusted R^2

Encouraged by the performance of the adjusted Bayesian measures in our simulations, we also compute estimates of trial-level surrogacy for the two ACCENT analyses that are adjusted for estimation error of the treatment effects. While it is certainly not surprising that the maximum likelihood estimates of R^2_{adj} continue to be numerically unavailable, we find the Bayesian posterior means of the adjusted measures to be noticeably higher than the unadjusted estimators for both older and newer ACCENT data sets.

For the older ACCENT trials, the posterior mean of R^2_{adj} is 0.932, seemingly indicating very good surrogacy. However, the standard deviation of the adjusted measure for these trials is 0.184, resulting in a posterior more variable than the posterior of the unadjusted measure. Thus, the increased surrogacy of TTR for OS in these trials suggested by R^2_{adj} comes with the caveat of less posterior confidence in such high surrogacy. Not surprisingly, a wide 95% credible interval for R^2_{adj} results, so that no strong conclusions can be made regarding the surrogacy of TTR for OS in the older ACCENT trials. The posterior probability that R^2_{adj} exceeds 0.90 is equal to 0.87, demonstrating that unusual "piling up" of the posterior for R^2_{adj} near 1 is occurring. Furthermore, concordance based on posteriors yielded by the adjusted model is low at 0.56, suggesting the trial-level model adjusting for estimation error of the treatment effects is worse than the unadjusted trial-level model at predicting unobserved treatment effects on the true endpoint, OS.

Turning to the newer ACCENT trials, the Bayesian adjusted surrogacy estimate is also increased from the unadjusted estimates, and to a considerably large extent. We find the posterior mean of R^2_{adj} is 0.935, again suggesting very high surrogacy of TTR for OS. This result seems counter-intuitive, given the apparent lack of association among the copula-estimated treatment effects for the new ACCENT trials in Figure 2.3. Clearly, the instability of R^2_{adj} we observed for a very low number of trials in simulations is still at play, which would lead us to a conclusion regarding the surrogacy of TTR for OS in these trials that is quite likely to be erroneous. Adding to this concern is a posterior probability that R^2_{adj} exceeds 0.90 which is equal to 0.873, reflecting that the posterior is "piling up" near 1 once again. Although perfect concordance of observed and predicted treatment effects on OS results from the adjusted trial-level model for the new ACCENT trials, a rather large posterior standard deviation of 0.179 yields 95% credible intervals that are likely too wide to be useful in determining trial-level surrogacy with accuracy.

2.5 Extension of the Bayesian Adjusted Model

Upon further exploration of the peculiar behavior of R^2_{adj}, we determined that the most influential component of the error-adjusted Bayesian model was not the choice of prior distributions, but the *scale* of the fixed trial-specific error covariances Ω_i. In particular, we demonstrate the impact of the scale of Ω_i on estimated R^2_{adj} through reconsideration of each ACCENT analysis, as follows.

Assume that the trial-specific error covariances in (2.7) are no longer fixed to their copula-stage estimates $\widehat{\Omega}_i$ as proposed by Burzykowski et al. (2001), but fixed to re-scaled estimates $\widehat{\Omega}_i/c$. On the precision scale, this corresponds to multiplying the trial-specific error precision Ω_i^{-1} by the same constant c. For each ACCENT analysis, we consider $c \in [1, 20]$ and display the posterior mean of R^2_{adj} as a function of c in the left panel of Figure 2.4.

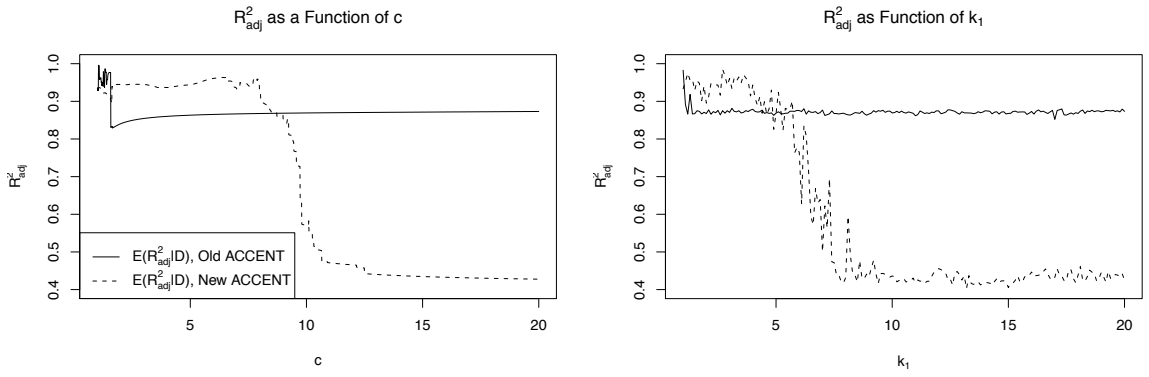

Figure 2.4: Left: Posterior means of R^2_{adj} as a function of multiplicative constant c for Ω_i^{-1}, $c \in [1, 20]$, old (—) and new (- - -) ACCENT trials. Right: Posterior means of R^2_{adj} as a function of lower bound k_1 of the uniform$(k_1, 100)$ prior for c, $k_1 \in [1, 20]$, old (—) and new (- - -) ACCENT trials.

When c is equal to 1, the usual scale for Ω_i proposed by Burzykowski et al. (2001) is obtained, and the estimates of R^2_{adj} correspond to the entries in Table 2.4. Here, by assuming $\Omega_i = \widehat{\Omega}_i$, we are fixing the variability of the error terms in (2.3) to be *equal* to the full variance-covariance of the treatment effect estimates approximated by maximum likelihood in the first-stage model. The high level of variability of the error terms forced by setting $c = 1$ may be too large. Furthermore, values of c less than 1 are implausible, as they represent cases in which the estimation error variance exceeds the total estimated variance of the treatment effects. Thus, we proceed by considering values of c that are greater than 1. By $c = 2$, the posterior mean of R^2_{adj} has returned (albeit rather abruptly) to a more plausible level of surrogacy for the older ACCENT trials. The new ACCENT trials, however, continue to yield inflated adjusted surrogacy estimates until approximately $c = 10$. At this point, one would assume the variability of estimation error is one-tenth the variability of the estimated treatment effects. Perhaps this smaller ratio is reasonable for the new ACCENT trials, where the total variability of each treatment effect estimate is estimated to be quite large (due to a small number of trials and low correlation

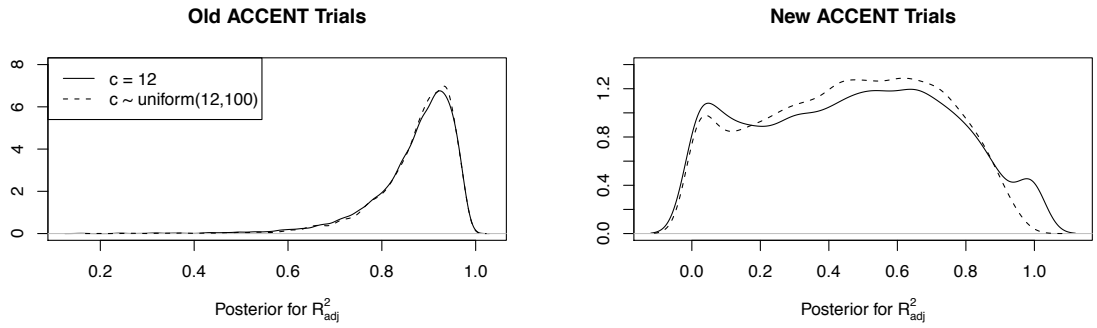

Figure 2.5: Posterior distributions of R^2_{adj} for fixed $c = 12$ (—) and $c \sim$uniform(12,100) priors (- - -) for the old and new ACCENT datasets.

of treatment effects across trials). As $c \to \infty$, the increasing error precisions $c\Omega_i^{-1}$, in combination with assumed zero mean error, forces the trial-specific error terms $(\epsilon_{a,i}, \epsilon_{b,i})$ to zero (estimation error is no longer present). It should come as no surprise, then, that the estimates of R^2_{adj} become *unadjusted* in this limit. Indeed, for $c = 20$, the posterior means of R^2_{adj} for each analysis are nearly equal to the posterior means of R^2_{un}.

It seems reasonable to conclude, then, that fixing c and thus Ω_i in the model for R^2_{adj} may be misleading, *especially* for blindly chosen values of c too close to 1. Thus, a logical next step in our Bayesian approach is to place a prior distribution on c, rather than choosing any single fixed value. If we specify $c \sim \text{uniform}(k_1, k_2)$, the choice of lower bound k_1 is undoubtedly more important than the choice of upper bound k_2, given what we learned in the left panel of Figure 2.4. We therefore proceed with this example by setting $k_2 = 100$, and observing R^2_{adj} estimates for each ACCENT analysis as k_1 varies from 1 to 20. These results are provided in the right panel of Figure 2.4.

By comparing R^2_{adj} estimates presented in each plot of Figure 2.4, we find that allowing c to vary through a prior distribution has the primary effect of contributing additional noise to surrogacy estimation. Indeed, the curves are generally quite

Table 2.5: Posterior means, standard errors, and 95% equal-tailed intervals for R^2_{adj} and $R^2_{adj,c}$ with $c = 12$ and $c \sim$uniform(12,100), for the old and new ACCENT trials.

Estimator		Mean	SE	95% Interval	Mean	SE	95% Interval
		9 Old ACCENT Trials			*6 New ACCENT Trials*		
$R^2_{adj}\|D$		0.932	0.184	(0.187, 0.999)	0.935	0.179	(0.213, 0.999)
$R^2_{adj,c}\|D$,	$c = 12$	0.870	0.090	(0.635, 0.972)	0.466	0.279	(0.006, 0.994)
$R^2_{adj,c}\|D$,	$c \sim$ uniform$(12, 100)$	0.875	0.084	(0.657, 0.972)	0.456	0.255	(0.008, 0.886)

similar; assigning a reasonable scale to the error covariances Ω_i, say by setting $c = 12$, yields R^2_{adj} estimates not unlike those obtained by choosing a lower bound of $k_1 = 12$ for the uniform prior on c. This suggests that placing a prior distribution on c may be unnecessary, so long as one has an idea of a reasonable scale on which to fix the error covariances. To confirm this, we plot the old and new ACCENT posterior distributions for R^2_{adj} with $c = 12$, overlaid with posterior distributions for R^2_{adj} where c was given a uniform$(12, 100)$ prior (Figure 2.5). Other than possibly eliminating some residual posterior probability near 1, the decision to place a prior on c makes little inferential difference for the new ACCENT analysis. Perhaps due to the small number of trials, or the weak correlation of estimated treatment effects across trials, no strong conclusions can be made regarding the error-adjusted surrogacy of TTR for OS. For the older ACCENT trials, the two posteriors are indistinguishable; giving c a prior distribution offers no advantages over simply choosing a reasonable scale for the error covariances relative to the data. Furthermore, we learn that the surrogacy of TTR for OS is quite likely to be high for these trials, given the observed data. This conclusion is consistent with existing knowledge and previous surrogacy analyses using these trials. These results are presented numerically in Table 2.5, where they may be compared with the naive estimates of R^2_{adj} presented in Section 2.4.

CHAPTER THREE

Bayesian Adaptive Trial Design for a Newly Validated Surrogate Endpoint

3.1 Introduction

In perhaps the most widely cited article in the surrogate endpoints literature, Prentice (1989) defined a surrogate endpoint as "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint." This definition, accompanied by a set of criteria proposed for the validation of such endpoints, inspired a wealth of subsequent work in this area as well as practical and theoretical debates which still persist today. Aside from its technical purpose, however, Prentice's definition reminds us that the ultimate goal of extensive surrogacy evaluation is the eventual use of a validated surrogate as the *primary* endpoint in a future trial.

While no single evaluative approach is universally accepted across all endpoint and disease types, a number of endpoints demonstrating consistently good surrogacy in particular settings have recently been identified and approved for primary use in future trials by regulatory authorities. Within the setting of adjuvant therapy in colon cancer, Sargent et al. (2005) demonstrated that disease-free survival (DFS) with 3 years median follow-up is a valid surrogate for 5 years median follow-up on overall survival (OS). This decision was not based on any single evaluative test of surrogacy, but on the potential surrogate's consistently good performance across an array of graphical and statistical checks that utilized patient-level historical data from multiple similar trials. In the setting of advanced colorectal cancer, progression-free survival was similarly validated as a surrogate for overall survival (Yothers, 2007). In the spirit of these recent developments, Shi and Sargent (2009) provided

a list of oncologic therapies which received regulatory approval based on trials using surrogate endpoints. As many authors in this area of literature have noted (see, for example, Burzykowski et al. (2001)), any validation process should include a formal assessment of the strength of association between treatment effects on the clinical and candidate surrogate endpoints across similar trials, in addition to simple within-trial correlative assessment of the endpoints themselves.

Even after a surrogate endpoint has been deemed "valid" for primary use in a future trial, however, a healthy skepticism *should* remain regarding the surrogate's ability to reflect the true treatment effect that would be observed given full follow-up on the clinical endpoint. Such caution is warranted for several reasons, including (1) agents for future testing may have different mechanisms of action; (2) improvement in outcomes outside of the specific treatments being tested are (one hopes) inevitable; and (3) each trial enrolls different patient mixes, which may impact disease natural history and/or relationships between endpoints. Though skepticism regarding a new primary endpoint seems reasonable for both clinical and regulatory purposes, few (if any) efforts have been made to formalize existing knowledge and uncertainty in the design of a future clinical trial where a validated surrogate takes on its intended role. This new trial should use the surrogate endpoint (which may be observed earlier or more often among patients) as the primary endpoint to establish efficacy, so that the benefits of having validated the endpoint may be reaped. At the same time, the fact that the new primary endpoint is really a surrogate, and perhaps only a recently-validated one, should not be ignored in the design of the trial.

We propose a trial design which adaptively checks accumulating data for consistency with relationships expected from the surrogacy validation process based on similar trials. At prospectively-defined checkpoints, we assess the performance of the surrogate and decide whether it should continue to be used to measure the effect of treatment. If the surrogate is behaving in the expected manner, we continue to

use it as the primary endpoint as the trial progresses to the next checkpoint; otherwise, we adaptively switch consideration back to the original clinical endpoint for the duration of the trial. Furthermore, at each checkpoint, we also check for early efficacy, inferiority, or futility of the experimental treatment arm compared to control, based on the primary endpoint at that time. If any of these conditions are met with adequate precision, we stop accrual, and thus avoid continuing a trial that is unnecessarily long, unnecessarily large, probably futile, or needlessly exposing many of its patients to a regimen which is likely to be inferior to a standard-of-care control. When faced with insufficiently precise information to establish early efficacy, inferiority, or futility, we continue accrual to the next prospectively-designed checkpoint. If a maximum number of patients are allowed to accrue and given reasonable follow-up, we compute a final measure of clinical benefit based on either a trusted surrogate or the original clinical endpoint.

The remainder of the chapter evolves as follows. In Section 3.2, we describe a Bayesian model for summarizing the relationship between treatment effects on the surrogate and clinical endpoints from multiple historical trials – the same relationship which previously validated the surrogate, and which we will reference during the new trial. We describe the design of this trial in Section 3.3, including its adaptive handling of efficacy, inferiority, futility, and expected surrogacy. Section 3.4 presents simulations for Type I error, power, and surrogacy discrimination compared to a standard O'Brien-Fleming approach (O'Brien and Fleming, 1979), while Section 3.5 demonstrates the new trial design using patient-level data from actual trials and a historically-validated endpoint in colon cancer (Sargent et al., 2005).

3.2 Bayesian Model for Historical Trials

Assume that a surrogate endpoint S has been validated to replace a primary clinical endpoint T based on a collection of historical trials indexed by $i \in \{1, ..., N\}$.

We further assume that S and T are time-to-event endpoints, as is often the case when a surrogate endpoint is desired, but the ensuing development is similar for endpoints of other types. Denote by Z the treatment assignment, where $Z = 1$ for experimental and $Z = 0$ for control. For historical trials containing more than one experimental arm, pairwise comparisons to the control arm may be required.

Many parametric and semi-parametric survival models exist, and for a given endpoint, only one such model may be appropriate. For our purposes, we assume both S and T follow Weibull(r, μ) distributions parameterized according to

$$f(x|r, \mu) = \mu r x^{r-1} \exp(-\mu x^r),$$

where $x > 0$ and $\mu, r > 0$. Further assuming unique baseline hazard functions for each trial and endpoint, the surrogate and clinical endpoints for patient $j \in \{1, ..., n_i\}$ from trial i are modeled as

$$S_{ij} \sim \text{Weibull}(r_i^S, \mu_{ij}^S) \tag{3.1}$$

$$\text{and} \quad T_{ij} \sim \text{Weibull}(r_i^T, \mu_{ij}^T), \tag{3.2}$$

where r_i^S and r_i^T are trial-specific shape parameters, and regressors z_{ij} and corresponding coefficients α_i and β_i are introduced through trial and patient-specific scale parameters $\mu_{ij}^S = \exp(\gamma_i^S + \alpha_i z_{ij})$ and $\mu_{ij}^T = \exp(\gamma_i^T + \beta_i z_{ij})$. Under this parameterization, median times until S and T for treatment arm Z in trial i are given by

$$\begin{aligned} m_{i,z}^S &= \{\log(2) \exp[-(\gamma_i^S + \alpha_i z_{ij})]\}^{1/r_i^S} \\ \text{and} \quad m_{i,z}^T &= \{\log(2) \exp[-(\gamma_i^T + \beta_i z_{ij})]\}^{1/r_i^T}. \end{aligned}$$

While joint modeling of S and T by copulas may be ideal for purposes such as surrogacy evaluation (Burzykowski et al., 2001), we choose marginal models to prevent unwanted learning between the endpoints (and treatment effects) within each trial.

Treatment effect estimates for each trial and endpoint may be obtained by maximum likelihood estimation, or alternatively, through Markov chain Monte Carlo (MCMC) estimation within a Bayesian framework (see Carlin and Louis (2009), Sec. 3.4).

Given that a linear theoretical relationship between α_i and β_i is reasonable, we assume

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_{aa} & \sigma_{ab} \\ \sigma_{ab} & \sigma_{bb} \end{bmatrix} \right), \tag{3.3}$$

which induces a regression line for α_i given β_i with intercept $\delta_0 = \alpha - \beta\sigma_{ab}/\sigma_{bb}$ and slope $\delta_1 = \sigma_{ab}/\sigma_{bb}$. Straightforward extensions are possible for treatment effect pairs (α_i, β_i) showing a nonlinear relationship. Using vague prior distributions on (α, β) and Σ, we obtain posterior distributions summarizing what is known about the relationship of treatment effects across the historical trials, quite possibly the same relationship that led to the validation of S as a surrogate endpoint for T.

3.3 Bayesian Adaptive Design

Given the model presented in Section 3.2 relating treatment effects on S and T across similar historical trials, we now describe the design of a future trial, indexed by $i = N + 1$, which uses the newly-validated surrogate endpoint S as the primary endpoint. In the context of a Phase III trial, we expect S to be the endpoint used to establish efficacy of the experimental treatment for regulatory purposes. This new trial will monitor the effect of treatment on both S and T, though in general, more events may be expected to occur (and occur earlier) for S than for T, hence the desirability of S as a surrogate for T. Any information accumulating for T will be used adaptively to assess the performance of the surrogate endpoint S throughout the trial, but will not influence the assessment of treatment effect on S.

Bayesian adaptive designs have been the subject of much recent research interest; see e.g., the textbook by Berry et al. (2010) for an overview and reference

list. In our trial, at various prospectively-defined interim check points, we check for consistency of accumulating information regarding the current trial's treatment effects, $(\alpha_{N+1}, \beta_{N+1})$, with their expected relationship based on (3.3). Given that the observed relationship between α_{N+1} and β_{N+1} at any time point is consistent with past experience, we proceed to check α_{N+1} for early efficacy and inferiority based on pre-specified posterior thresholds. If either reason for early stopping of the trial cannot be established, we check for futility of the trial based on the predictive probability of trial success (an efficacious treatment), before continuing accrual to the next interim checkpoint. If, on the other hand, the relationship between α_{N+1} and β_{N+1} is inconsistent with that given by (3.3), we adaptively switch to consideration of T (and β_{N+1}) for establishing early efficacy, inferiority, or futility. We assume up to K interim checks during the trial, which we index by $k = 1, ..., K$. The timing of checkpoints is generally associated with specific levels of accrual or percentages of observed (uncensored) events, so that stopping the trial early at any point may have the effect of decreasing overall sample size.

Though adaptive randomization could be considered (see Berry et al. (2010), Section 4.4), we maintain equal allocation to each of two treatment arms, and collectively denote the historical trials' data by D_h and the new trial's currently accrued patient data by D_k. We continue to assume that $S_{N+1,j}$ and $T_{N+1,j}$, $j = 1, ..., n_{N+1}$ follow Weibull distributions as in (3.1)-(3.2), but the same design may be easily implemented using other endpoint models. Enrolled patients not experiencing an event S or T by time k will be right-censored at their present length of follow-up for that endpoint. At a given checkpoint k, using vague priors and MCMC sampling to estimate the parameters of (3.1)-(3.2) for trial $i = N + 1$, we may obtain posterior distributions on the treatment effects α_{N+1} and β_{N+1} given current data D_k. In particular, for independent priors $\alpha_{N+1} \sim N(\mu_{0,\alpha}, \sigma^2_{0,\alpha})$, $\gamma^S_{N+1} \sim N(\mu_{0,\gamma_S}, \sigma^2_{0,\gamma_S})$, and $r^S_{N+1} \sim \text{Exp}(\lambda^S_0)$, the joint posterior for $(\alpha_{N+1}, \gamma^S_{N+1}, r^S_{N+1})$ given data D_k on

currently enrolled patients $j \in \{1, ..., n_{N+1}^k\}$ is proportional to the data likelihood times the joint prior,

$$\begin{aligned}\pi(\alpha_{N+1}, \gamma_{N+1}^S, r_{N+1}^S | D_k) &\propto L(\alpha_{N+1}, \gamma_{N+1}^S, r_{N+1}^S | D_k) p(\alpha_{N+1}) p(\gamma_{N+1}^S) p(r_{N+1}^S) \\ &\propto \prod_{j=1}^{n_{N+1}^k} \left\{ \left[\mu_{N+1,j}^S r_{N+1}^S (s_{N+1,j})^{r_{N+1}^S - 1} \right]^{\nu_j} \exp\left[-\mu_{N+1,j}^S (s_{N+1,j})^{r_{N+1}^S} \right] \right\} \\ &\quad \times \exp\left\{ -\lambda_0^S r_{N+1}^S - \frac{1}{2} \left[\left(\frac{\alpha_{N+1} - \mu_{0,\alpha}}{\sigma_{0,\alpha}} \right)^2 + \left(\frac{\gamma_{N+1}^S - \mu_{0,\gamma_S}}{\sigma_{0,\gamma_S}} \right)^2 \right] \right\}.\end{aligned}$$

Here, $\mu_{N+1,j}^S$ is defined as in (3.1), and ν_j is a censoring indicator equal to 1 if the jth individual has experienced event S prior to checkpoint k, and equal to 0 otherwise. The joint posterior $\pi(\beta_{N+1}, \gamma_{N+1}^T, r_{N+1}^T | D_k)$ associated with endpoint T may be derived similarly. For early interim checks, the variability in one or both of these posteriors may be large, and this uncertainty will be formally incorporated in any of the following adaptive decisions.

At a given checkpoint k, we adaptively assess the progress of trial $i = N + 1$ according to the following algorithm:

Algorithm 3.1 (Adaptive Design for a Surrogate Endpoint)

> Step 1. Surrogacy: Assess the consistency of $(\alpha_{N+1}, \beta_{N+1})$ with the historical relationship given by (3.3). Denote by $\pi(\alpha_{N+1}|D_k)$ the current posterior for the treatment effect on S in the new trial, and denote by $\pi(E(\alpha_{N+1}|\beta_{N+1})|D_h, D_k)$ the current posterior for the *expected* treatment effect on S, given both the effect on T in the new trial and historical uncertainty from (3.3). If these posteriors overlap in probability to a degree less than P_{surr}, discontinue consideration of the surrogate endpoint for the remainder of the trial, and perform Steps 2 through 4 using the treatment effect β_{N+1} for the endpoint T (skipping Step 1 at subsequent checkpoints). Otherwise, if $\pi(\alpha_{N+1}|D_k)$ and $\pi(E(\alpha_{N+1}|\beta_{N+1})|D_h, D_k)$ overlap to a degree greater than or equal to

P_{surr}, continue with S as the primary endpoint, and perform Steps 2 through 4 using α_{N+1}.

Step 2. Efficacy: Check for early efficacy by computing the posterior probability that α_{N+1} (or β_{N+1})) is less than 0, assuming an efficacious treatment extends the time to event endpoint. If $P(\alpha_{N+1} < 0|D_k) > P_{eff}$ (good surrogacy) or $P(\beta_{N+1} < 0|D_k) > P_{eff}$ (rejected surrogacy), where P_{eff} is some chosen threshold, stop the trial for early efficacy. Otherwise, continue to Step 3.

Step 3. Inferiority: Check for early inferiority by computing $P(\alpha_{N+1} > 0|D_k)$ (or $P(\beta_{N+1} > 0|D_k)$ under rejected surrogacy), assuming an inferior treatment *reduces* the time to event. If the posterior probability of inferiority of the experimental arm is greater than some threshold P_{inf}, stop the trial for early inferiority. Otherwise, continue to Step 4.

Step 4. Futility: Check for early futility by computing the empirical probability of trial success (the experimental treatment is efficacious) given full accrual to n_{N+1} and sufficient follow-up. In this case, eventual success is determined using a posterior predictive method: random samples from the joint posterior $\pi(\alpha_{N+1}, \beta_{N+1}, \gamma^S_{N+1}, \gamma^T_{N+1}, r^S_{N+1}, r^T_{N+1}|D_k)$ at the current checkpoint k are used to simulate event times for patients that would be observed if the trial were to continue to a pre-determined maximum length. If the updated posterior probability of efficacy after some minimum follow-up for the last hypothetical patient enrolled is greater than P_{eff}, the trial is considered a success. The empirical probability of trial success, computed over the random draws from $\pi(\alpha_{N+1}|D_k)$ (or $\pi(\beta_{N+1}|D_k)$), is compared to P_{fut}. The trial only continues to the next checkpoint k if this probability is greater than P_{fut}.

If the final interim checkpoint K is reached and the decision is made to continue enrollment, the trial proceeds to full accrual at the pre-set maximum sample size n_{N+1} determined by the sponsor. Generally, a final analysis of treatment benefit is performed after some minimum follow-up period on the last patient enrolled, so that some scientific contribution to the trial exists for every patient (Berry et al. (2010), p. 223). This final analysis consists of Steps 1-3 if the surrogate endpoint is still in use, or Steps 2-3 if the surrogate endpoint has been previously rejected. At the trial's end, we remain interested in both the surrogate endpoint's performance and effect of treatment after full accrual. The conclusions made with respect to the effect of the experimental treatment on final active outcome (S or T) will be those presented to regulatory authorities, assuming a Phase III confirmatory trial is being conducted. If the endpoint S is indeed a good surrogate for T, we might expect the information on the effect of treatment to be more precise for S than for T, as more observations on S were likely possible.

3.3.1 Checking Surrogacy

We now describe implementation of Step 1 from Algorithm 3.1 in more detail, for a given interim checkpoint k. Having obtained posterior distributions $\pi(\delta_1|D_h)$ and $\pi(\delta_0|D_h)$ for the slope and intercept, respectively, of the historical relationship between treatment effects on S and T (as described in Section 3.2), we compute the conditional posterior *we expect* for the current trial's treatment effect on S, given the current trial's treatment effect on T, denoted $\pi(E(\alpha_{N+1}|\beta_{N+1})|D_h, D_k)$. This may be computed using posterior samples from each of $\pi(\delta_1|D_h)$, $\pi(\delta_0|D_h)$, and $\pi(\beta_{N+1}|D_k)$, according to the historical fixed relationship $E(\alpha_i|\beta_i) = \delta_0 + \delta_1\beta_i$. Thus, to determine whether the behavior of the surrogate endpoint in the current trial is in agreement with historical trials, we approximate the amount of posterior

overlap by computing

$$2\min\{P(\alpha_{N+1} < \delta_0 + \delta_1\beta_{N+1}|D_h, D_k), P(\alpha_{N+1} > \delta_0 + \delta_1\beta_{N+1}|D_h, D_k)\}. \quad (3.4)$$

Assuming $\pi(\alpha_{N+1}|D_k)$ and $\pi(E(\alpha_{N+1}|\beta_{N+1})|D_h, D_k)$ are approximately symmetrically distributed, the multiplier 2 in (3.4) is required as roughly half of posterior samples within the region of overlap will be detected for a given inequality direction. Perfect surrogacy exists in those situations where the posterior means of α_{N+1} and $E(\alpha_{N+1}|\beta_{N+1})$ coincide; that is, when symmetry is present and $E(\alpha_{N+1}|D_k) = E[E(\alpha_{N+1}|\beta_{N+1})|D_h, D_k]$. If the empirical amount of overlap is greater than P_{surr}, we conclude based on available information that there is no reason to believe the surrogate is performing poorly; in other words, the current posterior distribution for α_{N+1} is centered along the regression line evaluated where posterior mass for β_{N+1} is currently located. Otherwise, if $\pi(\alpha_{N+1}|D_k)$ and $\pi(E(\alpha_{N+1}|\beta_{N+1})|D_h, D_k)$ overlap very little, we learn that the surrogate and former clinical endpoints are not relating in the historically-expected way in the current trial, and thus the surrogate endpoint should not be trusted. In general, we choose P_{surr} to be quite small (say, 0.10 or less), relying on the fact that the surrogate has been previously validated and rejection of a surrogate mid-trial should be difficult. Lingering uncertainty regarding the validating relationship will manifest itself in wider posteriors for δ_0 and δ_1, which will also serve to make a surrogate more difficult to reject. When dissimilar posterior distribution shapes are present, a measure of shared posterior information more formal than (3.4) may be required. For example, Kullback-Leibler divergence (Kullback and Leibler, 1951) may be used to estimate the information shared by $\pi(\alpha_{N+1}|D_k)$ and $\pi(E(\alpha_{N+1}|\beta_{N+1})|D_h, D_k)$, either based on closed forms if available, or analytical approximations otherwise.

Although this design combines past and present sources of uncertainty in order to evaluate surrogacy, it is important to note that it also explicitly (and correctly) prohibits Bayesian learning between the collection of historical trials indexed by

$i \in \{1, ..., N\}$ and the current trial, indexed by $i = N + 1$. Prior to the new trial, information is shared across similar historical trials in estimating parameters of (3.3), or equivalently, in confirming the trial-level surrogacy which previously validated S for T. This historical snapshot of the surrogacy relationship across trials, as well as posterior uncertainty for the slope δ_1 and the intercept δ_0 that summarize this relationship, are considered fixed prior to and during the new trial that uses S as a primary endpoint. In other words, data accumulated during the new trial do not inform or update the posteriors from the historical model (3.3), as this would undermine interim checks for surrogacy. Likewise, posterior information from fitting (3.3) to historical treatment effects is not used to inform estimation of the treatment effects α_{N+1} and β_{N+1} in the new trial. Thus, while we adaptively monitor the surrogacy of S for T in the new trial for consistency with historical trials, any conclusions made with respect to the intervention (efficacy, inferiority, or futility) may still be regarded as independent from historical information for regulatory purposes.

3.4 Simulation Study

As is generally the case for trials designed with Bayesian adaptive stopping rules, frequentist operating characteristics such as Type I error and power are most conveniently assessed through simulation studies. We note that our simulation settings were intentionally chosen to reflect characteristics of actual trials presented in the data analysis of Section 3.5. Throughout, we assume *fixed* historical validation data in the form of treatment effects (α_i, β_i), $i = 1, ..., 25$ arising from model (3.3) with mean $(\alpha, \beta) = (-0.5, -0.5)$, variances $\sigma_{aa} = \sigma_{bb} = 1$, and covariance $\sigma_{ab} = 0.95$. These specifications correspond to a historical slope of 1, historical intercept 0, and historical correlation across treatment effect pairs equal to 0.95. We assume this strong relationship previously played a role in validating S as a surrogate for T,

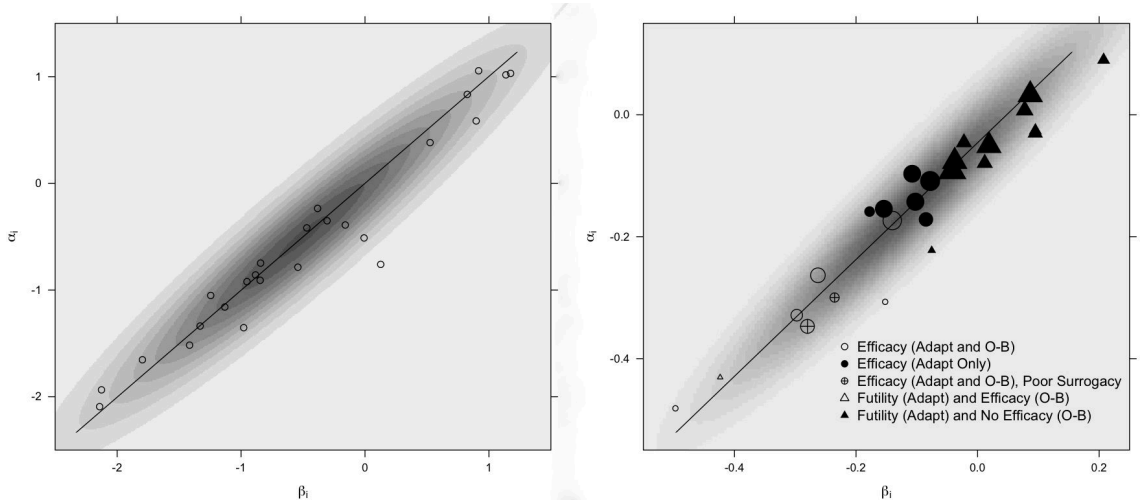

Figure 3.1: Scatterplots of historical treatment effect pairs (α_i, β_i), $i \in \{1, ..., 25\}$, for the simulated historical trials (left) and the ACCENT trials (right). The simulated effects are superimposed on an image plot of the bivariate normal density from which they were sampled, while the ACCENT effects are superimposed on an image plot of the bivariate normal distribution estimated from model (3.3) and weighted by the square root of trial size. Regression lines based on $E(\delta_0|D_h)$ and $E(\delta_1|D_h)$ from model (3.3) fits are included for each case, where D_h collectively denotes data from the historical trials. The position of each ACCENT treatment effect pair is based on the original trials with all available patient follow-up, while the symbol for each treatment effect pair denotes the conclusions that would have been reached had the trial been performed according to our adaptive design or an O'Brien-Fleming design.

perhaps by a meta-analytic surrogacy measure such as R^2_{trial}, which is theoretically equal to 0.9025 in this case (see Burzykowski et al. (2001) for a discussion of R^2_{trial}). A scatterplot of simulated treatment effect pairs from the historical trials, (α_i, β_i), superimposed on a contour plot of their underlying bivariate normal distribution, is presented in the left window of Figure 3.1.

After generating 25 random treatment effect pairs (α_i, β_i) from (3.3) as specified above, we estimate the parameters of the same model – which we now assume to be unknown – using MCMC with vague prior distributions. In particular, we place $N(0, \tau = 0.0001)$ priors on each of the mean parameters (where τ is the precision, or reciprocal of the variance), and we choose a Wishart$(2, M)$ prior for the *precision* matrix Σ^{-1}, where M is a diagonal matrix with $diag(M) = (10^{-6}, 10^{-6})$. We acquire

posterior sample chains of length 10,000 for each parameter, after a burn-in of 1,000 iterations. With these samples, we derive posteriors for the slope δ_1 and intercept δ_0 of the regression line induced by model (3.3), which remain fixed throughout our simulations and will be referenced as we assess surrogacy in the new trial.

3.4.1 Data Generation and Settings

While the posterior uncertainty for δ_0 and δ_1 derived from fitting model (3.3) remains fixed throughout our simulations, we vary some characteristics of the new trial in order to assess the operating characteristics of its design. We continue to assume that S and T follow Weibull distributions parameterized as in (3.1)-(3.2), and consider a maximum sample size of $n = 2000$ patients with equal allocation to two treatment arms. Throughout, we fix $r^S_{N+1} = r^T_{N+1} = 2$, but common shape parameters need not be assumed; indeed, we relax this assumption in the data analysis of Section 3.5. At the top level of the simulation, we specify true median event times $m^S_{N+1,0}$ and $m^T_{N+1,0}$ associated with the control group ($Z = 0$) for each endpoint in the new trial, as well as true improvement ratios of the experimental over control groups (in terms of median event times) for each endpoint. We denote these true improvement ratios by Δ_S and Δ_T. Together, the specified shape parameters, baseline median event times, and improvement ratios determine the true underlying Weibull regression parameters according to the following equations:

$$\gamma^S_{N+1} = -\log[(m^S_{N+1,0})^{r^S_{N+1}}/\log(2)] \qquad \alpha_{N+1} = -\log[(\Delta_S)^{r^S_{N+1}}]$$
$$\gamma^T_{N+1} = -\log[(m^T_{N+1,0})^{r^T_{N+1}}/\log(2)] \qquad \beta_{N+1} = -\log[(\Delta_T)^{r^T_{N+1}}].$$

These in turn determine the true scale parameters (regression components) given by $\mu^S_{N+1,j} = \exp(\gamma^S_{N+1} + \alpha_{N+1} z_{N+1,j})$ and $\mu^T_{N+1,j} = \exp(\gamma^T_{N+1} + \beta_{N+1} z_{N+1,j})$ for $j = 1, ..., n_{N+1}$.

For each scenario and endpoint, we generate random event times for patient j in trial $i = N+1$ by $S_{N+1,j} \sim \text{Weibull}(r^S_{N+1}, \mu^S_{N+1,j})$ and $T_{N+1,j} \sim \text{Weibull}(r^T_{N+1}, \mu^T_{N+1,j})$,

with shape and scale parameters fixed at their true values. As an alternative, one could first generate parameter values from highly informative "design priors" centered at the true values, and use these in turn to generate random Weibull data (see, for example, Berry et al. (2010), p. 72). Motivated by the assumed prior validation of S as a surrogate for T, we only consider the likely situation where S is known to have an earlier median event time than T in the control arm. Specifically, we choose $m^S_{N+1,0} = 350$ days and $m^T_{N+1,0} = 700$ days.

At this stage, we also generate random accrual dates for each patient according to a discrete uniform$(1, b)$ distribution, where b is chosen to be twice the true median event time for T in the control group. While this choice of b may seem arbitrary, it ensures that both early stopping rules and the use of a surrogate remain useful. Accrual dates are assumed unrelated to treatment assignment, so at any interim checkpoint k, we expect approximately equal enrollment in each treatment arm. Throughout the simulation studies, we define 3 interim checkpoints to be the dates on which 25%, 50%, and 75% of patients have experienced events. Trials which continue beyond the third checkpoint have final analyses after 100% of patients have experienced events. In the data analysis of Section 3.5, timing of checkpoints and final analyses will be adapted for trials with a high level of censoring throughout.

With the Weibull shape parameters and true median event times fixed in each simulation, we focus primarily on the effect of varying the improvement ratios Δ_S and Δ_T. When $\Delta_S = \Delta_T = 1$, we have good surrogacy in the new trial, but no improvement in the experimental arm over the control arm (Type I error case). When $\Delta_S = \Delta_T \neq 1$, we continue to have good surrogacy, but now with efficacy or inferiority of the experimental treatment (power case). In cases where $\Delta_S \neq \Delta_T$, the treatment effects α_{N+1} and β_{N+1} will be discordant, no longer falling along the regression line induced by model (3.3), indicating poor surrogacy. We study a range of cases where Δ_S and Δ_T agree, representing the trial's operating characteristics

assuming good surrogacy of S for T. Similarly, we consider cases where Δ_S and Δ_T disagree, and where reliance on S to determine a treatment's benefit could lead to an erroneous conclusion.

We perform $R = 100$ replications (hypothetical trials) for each scenario, which are described in Table 3.1. At each checkpoint k reached within a simulated trial, we estimate the parameters of models (3.1) and (3.2) using available data on currently accrued patients. If the endpoint S was not already rejected at an earlier checkpoint, we check check its surrogacy for T and decide to continue trusting S or switch consideration to T. At the same checkpoint, we may also stop the trial for early efficacy, early inferiority, or likely futility based on the current endpoint. If none of the stopping rules are met, we continue accrual to the next prospectively-designed checkpoint. Throughout, these decisions are based on rather conservative posterior thresholds given by $P_{surr} = 0.01$, $P_{eff} = 0.99$, $P_{inf} = 0.99$, and $P_{fut} = 0.05$, though one could choose less conservative bounds as the trial progresses. For regulatory purposes, we use $N(0, 0.0001)$ priors on the treatment effects α_{N+1} and β_{N+1}, but choose more informative priors for δ^S and δ^T (centered at true values) to focus our estimative power on the parameters of interest. In each case, we obtain posterior samples of length 10,000 after 1,000 burn-in iterations.

To facilitate comparison of our design against a commonly-used frequentist design with interim analyses, we perform parallel analyses for early efficacy based on log-rank tests with O'Brien-Fleming stopping rules, where the same set of checkpoints are used and an overall $\alpha = 0.05$ is maintained for each trial. We note that our choice of Bayesian efficacy threshold, $P_{eff} = 0.99$, is more conservative than an overall $\alpha = 0.05$ threshold when considered across four potential sequential analyses per trial.

Table 3.1. Scenarios explored in the simulation study.

Good Surrogacy	Poor Surrogacy
No effect on S or T:	Positive effect on T, no effect on S:
$\Delta_S = \Delta_T = 1.0$	$\Delta_S = 1$, $\Delta_T = 1.1$
	$\Delta_S = 1$, $\Delta_T = 1.2$
Equal effects on S and T:	$\Delta_S = 1$, $\Delta_T = 1.4$
$\Delta_S = \Delta_T = 0.8$	
$\Delta_S = \Delta_T = 0.9$	Positive effect on S, no effect on T:
$\Delta_S = \Delta_T = 1.1$	$\Delta_S = 1.1$, $\Delta_T = 1$
$\Delta_S = \Delta_T = 1.2$	$\Delta_S = 1.2$, $\Delta_T = 1$
$\Delta_S = \Delta_T = 1.4$	$\Delta_S = 1.4$, $\Delta_T = 1$

3.4.2 Simulation Results

In Tables 3.2 and 3.3, we present the results of the simulation scenarios outlined in Table 3.1 where good surrogacy is assumed and the true treatment effect is varied from strongly inferior to strongly efficacious. In Tables 3.4 and 3.5, we present results for scenarios with poor surrogacy, with a treatment effect on T not reflected by a treatment effect on S, and a treatment effect on S not reflected by a treatment effect on T, respectively.

Simulation results for scenarios with good surrogacy ($\Delta_S = \Delta_T = \Delta$) are presented in Tables 3.2 and 3.3, where we consider hypothetical trials with treatment worse than control ($\Delta = 0.8, 0.9$), equal to control ($\Delta = 1$), and better than control ($\Delta = 1.1, 1.2, 1.4$) in terms of improved median S and T. Bearing in mind the order in which the steps of Algorithm 3.1 are performed at each checkpoint, we find that most simulated trials are stopped early for the correct decision. When $\Delta = 0.80$, the median event times for S and T in the treatment group are 80% of the median times in the control group, and all 100 simulated trials stop at the first interim checkpoint for treatment inferiority. Once median event times in the treatment group increase to 90% of the median event times in the control group, fewer trials are stopped for inferiority and more are stopped for futility. In each inferiority case,

Table 3.2: Simulation results for scenarios with good surrogacy ($\Delta_S = \Delta_T = \Delta$) and either inferiority or no treatment effect. Mean(SS) is the mean sample size across hypothetical trials. P(1), P(2), P(3), and P(Full) are the proportions of trials showing poor surrogacy or that were stopped at interim checkpoints 1, 2, 3, or run to full follow-up, respectively. The final column reports total percentages for poor surrogacy and each stopping reason across time points, by endpoint.

Δ	Mean(SS)	Reason	Endpoint	P(1)	P(2)	P(3)	P(Full)	Total
0.8	975	*Surrogacy*		-	-	-	-	*0.00*
		Efficacy	S	-	-	-	-	-
			T	-	-	-	-	-
		Inferiority	S	1.00	-	-	-	1.00
			T	-	-	-	-	-
		Futility	S	-	-	-	-	-
			T	-	-	-	-	-
	2000	Efficacy (O-F)	S	-	-	-	-	0.00
0.9	999	*Surrogacy*		-	-	-	-	*0.00*
		Efficacy	S	-	-	-	-	-
			T	-	-	-	-	-
		Inferiority	S	0.61	-	-	-	0.61
			T	-	-	-	-	-
		Futility	S	0.39	-	-	-	0.39
			T	-	-	-	-	-
	2000	Efficacy (O-F)	S	-	-	-	-	0.00
1.0	1259	*Surrogacy*		*0.02*	-	-	-	*0.02*
		Efficacy	S	-	0.01	-	-	0.01
			T	0.01	-	-	-	0.01
		Inferiority	S	-	-	-	-	-
			T	-	-	-	-	-
		Futility	S	0.61	0.28	0.07	0.01	0.97
			T	0.01	-	-	-	0.01
	2000	Efficacy (O-F)	S	-	-	-	0.02	0.02

no trials demonstrate efficacy according to the O'Brien-Fleming approach. When the treatment arm has no benefit ($\Delta = 1$), 98% of trials are stopped early for futility, with most stopping at the first checkpoint. In this case, the trials that incorrectly stopped for efficacy or inferiority together yield an estimated Type I error rate of 2%, the same error rate observed for the O'Brien-Fleming approach. Once we set Δ to be greater than 1 in Table 3.3, yielding improved median event times for the experimental treatment relative to control (and thus representing power calculations), the vast majority of trials are correctly stopped early for treatment

Table 3.3: Simulation results for scenarios with both good surrogacy ($\Delta_S = \Delta_T = \Delta$) and efficacy. Mean(SS) is the mean sample size across hypothetical trials. P(1), P(2), P(3), and P(Full) are the proportions of trials showing poor surrogacy or that were stopped at interim checkpoints 1, 2, 3, or run to full follow-up, respectively. The final column reports total percentages for poor surrogacy and each stopping reason across time points, by endpoint.

Δ	Mean(SS)	Reason	Endpoint	P(1)	P(2)	P(3)	P(Full)	Total
1.1	1380	*Surrogacy*		-	-	-	-	*0.00*
		Efficacy	S	0.48	0.33	0.11	0.05	0.97
			T	-	-	-	-	-
		Inferiority	S	-	-	-	-	-
			T	-	-	-	-	-
		Futility	S	0.02	-	-	0.01	0.03
			T	-	-	-	-	-
	1759	Efficacy (O-F)	S	-	0.55	0.36	0.09	1.00
1.2	1073	*Surrogacy*		*0.01*	-	-	-	*0.01*
		Efficacy	S	0.99	-	-	-	0.99
			T	0.01	-	-	-	0.01
		Inferiority	S	-	-	-	-	-
			T	-	-	-	-	-
		Futility	S	-	-	-	-	-
			T	-	-	-	-	-
	1316	Efficacy (O-F)	S	0.52	0.48	-	-	1.00
1.4	1110	*Surrogacy*		-	-	-	-	*0.00*
		Efficacy	S	1.00	-	-	-	1.00
			T	-	-	-	-	-
		Inferiority	S	-	-	-	-	-
			T	-	-	-	-	-
		Futility	S	-	-	-	-	-
			T	-	-	-	-	-
	1110	Efficacy (O-F)	S	1.00	-	-	-	1.00

efficacy under each approach. However, our adaptive design tends to stop a larger percentage of trials for efficacy at earlier checkpoints compared to the O'Brien-Fleming design, thus offering significant savings in time and cost. In all adaptive cases, we obtain an average reduction in sample size of more than 31%, while many O'Brien-Fleming cases show no savings in sample size.

Scenarios with poor surrogacy ($\Delta_S \neq \Delta_T$) also show promising results, which we present in Tables 3.4 and 3.5. The results in Table 3.4 represent trials where a beneficial effect on the former clinical endpoint T is not reflected in a beneficial

Table 3.4: Simulation results for scenarios with poor surrogacy ($\Delta_S \neq \Delta_T$) and a greater treatment effect on T. Mean(SS) is the mean sample size across hypothetical trials. P(1), P(2), P(3), and P(Full) are the proportions of trials showing poor surrogacy or that were stopped at interim checkpoints 1, 2, 3, or run to full follow-up, respectively. The final column reports percentages for poor surrogacy and each stopping reason across time points, by endpoint.

Δ_S	Δ_T	Mean(SS)	Reason	Endpoint	P(1)	P(2)	P(3)	P(Full)	Total
1	1.1	1299	*Surrogacy*		*0.06*	*0.03*	-	-	*0.09*
			Efficacy	S	0.01	-	0.02	0.01	0.04
				T	0.06	0.03	-	-	0.09
			Inferiority	S	-	-	-	-	-
				T	-	-	-	-	-
			Futility	S	0.51	0.26	0.10	-	0.87
				T	-	-	-	-	-
		2000	Efficacy (O-F)	S	-	-	0.02	0.01	0.03
1	1.2	1284	*Surrogacy*		*0.25*	*0.10*	*0.02*	-	*0.37*
			Efficacy	S	0.01	-	-	-	0.01
				T	0.25	0.10	0.02	-	0.37
			Inferiority	S	-	-	-	-	-
				T	-	-	-	-	-
			Futility	S	0.34	0.18	0.10	-	0.62
				T	-	-	-	-	-
		2000	Efficacy (O-F)	S	-	-	0.01	0.02	0.03
1	1.4	1073	*Surrogacy*		*0.86*	*0.10*	-	-	*0.96*
			Efficacy	S	-	-	-	-	-
				T	0.86	0.10	-	-	0.96
			Inferiority	S	-	-	-	-	-
				T	-	-	-	-	-
			Futility	S	0.04	-	-	-	0.04
				T	-	-	-	-	-
		2000	Efficacy (O-F)	S	-	-	0.01	0.02	0.03

treatment effect on the surrogate endpoint S. These may be the scenarios of greatest interest (or concern) in practice, as we usually anticipate improved experimental outcomes over control when designing a trial, corresponding to values of Δ_T greater than 1. It may be feared that the surrogate endpoint may fail to detect a truly beneficial effect of treatment, as indicated by concurrently setting $\Delta_S = 1$. However, poor surrogacy is often detected among these cases, and detected earlier as the treatment effects on S and T become more discordant. This is indeed comforting, as we would like to avoid trusting an invalid surrogate well into an expensive and

Table 3.5: Simulation results for scenarios with poor surrogacy ($\Delta_S \neq \Delta_T$) and a greater treatment effect on S. Mean(SS) is the mean sample size across hypothetical trials. P(1), P(2), P(3), and P(Full) are the proportions of trials showing poor surrogacy or that were stopped at interim checkpoints 1, 2, 3, or run to full follow-up, respectively. The final column reports percentages for poor surrogacy and each stopping reason across time points, by endpoint.

Δ_S	Δ_T	Mean(SS)	Reason	Endpoint	P(1)	P(2)	P(3)	P(Full)	Total
1.1	1	1377	*Surrogacy*		*0.08*	*0.02*	-	-	*0.10*
			Efficacy	S	0.44	0.27	0.16	0.03	0.90
				T	-	-	-	-	-
			Inferiority	S	-	-	-	-	-
				T	0.01	0.01	-	-	0.02
			Futility	S	-	-	-	-	-
				T	0.07	0.01	-	-	0.08
		1752	Efficacy (O-F)	S	-	0.57	0.36	0.06	0.99
1.2	1	1091	*Surrogacy*		*0.31*	-	-	-	*0.31*
			Efficacy	S	0.66	0.03	-	-	0.69
				T	-	-	-	-	-
			Inferiority	S	-	-	-	-	-
				T	-	-	-	-	-
			Futility	S	-	-	-	-	-
				T	0.30	0.01	-	-	0.31
		1336	Efficacy (O-F)	S	0.48	0.52	-	-	1.00
1.4	1	1382	*Surrogacy*		*0.97*	-	-	-	*0.97*
			Efficacy	S	0.03	-	-	-	0.03
				T	0.01	0.02	0.02	0.01	0.06
			Inferiority	S	-	-	-	-	-
				T	-	-	-	-	-
			Futility	S	-	-	-	-	-
				T	0.58	0.18	0.12	0.03	0.91
		1119	Efficacy (O-F)	S	1.00	-	-	-	1.00

lengthy trial. It is interesting to note that many trials stop for futility of S in less discordant cases, particularly when $\Delta_S = 1$ and $\Delta_T = 1.1$ or 1.2; this is due to the fact that futility is determined through S using predictive probabilities based on $\pi(\alpha_{N+1}|D_k)$. With $\Delta_S = 1$, we would expect to reach a futility decision, given that surrogacy is not so poor in Step 1 as to preempt Step 4 of Algorithm 3.1. The O'Brien-Fleming approach with S as its primary endpoint is unable to detect poor surrogacy, and continues to trust S throughout the trial. In most cases, this design determines that the experimental treatment is ineffective, often late in the

trial. In reality, S is unable to measure the true positive effect on T, and thus the O'Brien-Fleming trials here may prevent regulatory acceptance of truly beneficial treatments.

For symmetry, we also consider cases where a positive effect on S ($\Delta_S > 1$) is not supported by a positive effect on T ($\Delta_T = 1$), with results presented in Table 3.5. Although simulated trials stop more frequently for early efficacy based on S than for futility based on T in the cases of $\Delta_S = 1.1$ or 1.2 and $\Delta_T = 1$, this is not entirely surprising, as determinations of efficacy are based on S through $\pi(\alpha_{N+1}|D_k)$ when surrogacy is not so bad as to switch consideration to T. Once the differences in treatment effect on S and T become more pronounced, trials are correctly stopped early for futility based on T at a greater rate. The O'Brien-Fleming approach based on S incorrectly stops for efficacy in nearly all cases where S shows a positive effect and T does not, representing situations where regulatory acceptance might be awarded to a truly ineffective treatment. Across all cases with poor surrogacy, we obtain an average reduction in total sample size of more than 31%, while many O'Brien-Fleming cases show no savings in sample size.

3.5 Example: Adaptive Monitoring of ACCENT Trials

In this section, we illustrate our new design with an example from colorectal cancer, the third most common cancer in the United States with approximately 145,000 new cases diagnosed every year (Sargent et al., 2005). When no interventions are administered to patients with node-positive disease after primary resection, approximately half will experience relapse and eventually die as a result of their disease. Sargent et al. (2005) demonstrated that disease-free survival (DFS) with a median of 3 years follow-up is a valid surrogate for overall survival (OS) with a median of 5 years follow-up in the adjuvant setting, based on a variety of graphical and statistical approaches. The trials used in the validation process were provided

by the Adjuvant Colon Cancer End Points (ACCENT) Group, and contain individual patient data from 18 randomized phase II and phase III trials for adjuvant therapy in colon cancer. These trials were conducted from 1977 to 1999, and collectively include 20,898 patients assigned to 43 treatment arms, composed of 34 active treatment arms (with at least one fluorouracil (FU)-based chemotherapy arm per trial) and 9 surgery-only arms. Using patient-level data from the ACCENT trials and maintaining the natural ordering of the treatment arms in the original trial designs, we consider pairwise comparisons of experimental to control arms for a total of $N = 25$ trial units.

The ACCENT trials are ideally suited for re-evaluation in our adaptive design framework for a number of reasons, in addition to the well-established good surrogacy of DFS for OS in this setting. First, the median observed DFS time across trials and treatment groups is 643 days, compared to a median OS time of 1093 days, a difference of well over 1 year. As long as the effect of treatment on DFS is a good predictor of the effect of treatment on OS, the amount of follow-up required to observe an adequate number of events will be substantially less for the surrogate endpoint. Second, accrual is relatively slow, ranging from 17 patients to 126 patients per month to fully enroll trials ranging in size from 200 to over 2000 individuals. Third, the maximum follow-up within trials is usually well over 10 years, implying substantial ongoing effort and operational expense.

Considering all available patient data without imposing additional censoring to construct specific median follow-up times (as in Sargent et al. (2005)), we use the 25 trial-level units as historical trials in our design, which collectively indicate that DFS is a valid endpoint to use as a surrogate for OS in a future adjuvant colon cancer trial. A scatterplot of estimated treatment effects $(\hat{\alpha}_i, \hat{\beta}_i)$ from models (3.1) and (3.2) is shown in the right panel of Figure 3.1, superimposed on an image plot of the bivariate Normal distribution estimated from model (3.3). The treatment

effects on OS and DFS are highly linearly related, with Pearson correlation 0.9541 and estimated trial-level surrogacy $\hat{R}^2_{trial} = 0.9103$. This strong relationship across similar trials previously helped to justify the surrogacy of DFS for OS (Sargent et al., 2005).

In order to demonstrate how our design performs for realistic "future" trials, we will imagine that each of the ACCENT trials in turn has yet to be conducted until now. We begin by fitting models (3.1) and (3.2) to the full data from each trial, obtaining estimates of the treatment effect pairs (α_i, β_i) for $i \in \{1, ..., 25\}$. At this stage, MCMC estimation was used with diffuse normal priors for regression parameters and vague exponential priors for shape parameters. In the second stage of estimation, we fit model (3.3) to the estimated treatment effects from the first stage, using diffuse normal priors for the mean components (α, β) and a vague Wishart prior for the precision matrix Σ^{-1}. This in turn yields posteriors $\pi(\delta_1|D_h)$ and $\pi(\delta_0|D_h)$ for the historical slope and intercept, respectively, summarizing historical knowledge and uncertainty for the ACCENT trials' relationship in the right panel of Figure 3.1.

Now, with this historical picture fixed, we effectively re-run each ACCENT trial under our new adaptive design. We use the actual dates of enrollment and event times for each patient within each trial, and assume trial sizes were chosen to observe uncensored outcomes for a total of 25% of patients by the end of each trial. Thus, within each trial, we perform interim checks on those dates where 25%, 50%, and 75% of the total desired number of events have occurred. For each trial and checkpoint, we choose $P_{surr} = 0.10$, $P_{fut} = 0.05$, and $P_{eff} = P_{inf} = 0.95$. If the decision is made at the final interim checkpoint to continue, we follow all patients until a 25% event rate is observed before performing a final analysis. Estimation of models (3.1) and (3.2) at each checkpoint is based only on those patients actually enrolled, and is implemented using vague priors on all parameters, with MCMC chains of length

100,000 after 10,000 burn-in samples. We present the poor surrogacy and stopping frequencies across the 25 trials at each interim point, by reason, in Table 3.6.

Table 3.6: Frequencies at which poor surrogacy (based on DFS for OS), efficacy, inferiority, or futility were determined at each time point (percentages of the maximum number of desired events for final analysis) for the 25 ACCENT trials, by endpoint. The total number of ACCENT trials stopped at each time point are also given for each design (adaptive and O'Brien-Fleming). Trials with futility results at the final checkpoint had insufficient evidence to reject the null hypothesis of no treatment effect after the desired number of events had been observed.

	Endpoint	25% Max Events	50% Max Events	75% Max Events	100% Max Events	Total
Surrogacy		*2*	*0*	*0*	*0*	*2 (8%)*
Efficacy	S	5	4	2	0	11 (44%)
	T	2	0	0	0	2 (8%)
Inferiority	S	0	0	0	0	0 (0%)
	T	0	0	0	0	0 (0%)
Futility	S	4	1	0	7	12 (48%)
	T	0	0	0	0	0 (0%)
Adaptive Design Totals	S, T	11 (44%)	5 (20%)	2 (8%)	7 (28%)	25 (100%)
O'Brien-Fleming Totals	S	0	4 (16%)	3 (12%)	1 (4%)	8 (32%)

Among the 25 ACCENT trials considered in our new adaptive framework, 13 trials stopped early for efficacy of S with good surrogacy or efficacy of T with poor surrogacy. Of these, 7 trials stopped after the first interim check. Given the original treatment effect estimates based on all available follow-up in Figure 3.1, this is not surprising, as most plotted estimates $\hat{\alpha}_i$ are less than 0. Those ACCENT trials demonstrating efficacy at any time point within our adaptive design have original treatment effect pairs denoted by a circle in Figure 3.1. Filled circles denote trials where the O'Brien-Fleming design was not able to determine the positive treatment effect, while unfilled circles represent trials where both approaches measured a treatment benefit. A special case of these trials, where efficacy was established by both approaches but through T for the adaptive approach, is given by an unfilled circle surrounding a +. In the absence of an adaptive procedure, a decision based on a poor surrogate would have been made in these two trials.

ACCENT trials demonstrating futility at any time according to the adaptive design have original treatment effect pairs denoted by a triangle in Figure 3.1. It is interesting to note that most of the original treatment effects for these trials are located near the null treatment effect point $(\alpha_i, \beta_i) = (0, 0)$; the conclusions reached in the adaptive setting are similar to those reached in the non-adaptive historical setting, but with less follow-up required. Of the 12 futile trials identified by the adaptive approach, only one very small trial showed a treatment benefit under O'Brien-Fleming approach, which is represented by an unfilled triangle. The other O'Brien-Fleming trials failed to detect efficacy, thus agreeing with the adaptive approach, and are shown as filled triangles. We find that none of the adaptively-designed ACCENT trials reach a decision of treatment inferiority compared to control. It may seem surprising that the original treatment effect pairs for the two trials showing poor surrogacy fall somewhat close to the line indicating perfect surrogacy, but it should be noted that both trials are relatively large in size with strong positive treatment effects. The high precision of $\pi(\alpha_{N+1}|D_k)$ in each case may have prevented substantial overlap with $\pi(E(\alpha_{N+1}|\beta_{N+1})|D_h, D_k)$. However, neither case is costly in terms of wrong conclusions or additional resources, as the correct treatment benefit was observed (only in terms of T) at the same time that surrogacy was rejected. Overall, we found an average percent reduction in total sample size from the original trials of 11.8% (range: 0% to 69%) for our adaptive approach, but only 0.08% (range: 0% to 2%) for the O'Brien-Fleming approach, indicating that substantial time and resources could have been saved in most trials under our design.

CHAPTER FOUR

Flowgraph Modeling of Disease Progression for Evaluating Surrogate Endpoints

4.1 Introduction

Flowgraphs first appeared in the engineering and reliability literature (see, e.g., Dorf (1989) and Whitehouse (1983)), and statistical modeling techniques were later introduced to the statistics and probability disciplines by Butler and Huzurbazar (1997). Developed to be more flexible and realistic than traditional approaches to multi-state modeling, flowgraphs do not depend on Markov assumptions or exponential waiting times, are computationally inexpensive, and handle censoring and incomplete data easily and effectively (Butler and Huzurbazar, 1997). Furthermore, vastly different waiting time distributions may be assigned to different transitions within a system, and unique waiting time distributions may depend on covariates. Assumptions for flowgraph modeling are minimal: the probabilities associated with possible transitions out of each state must add to 1, and moment generating functions for waiting times between states must exist. Competing risks and conditional events are automatically incorporated into the structure of a flowgraph, and hazard functions associated with any composite or overall times-to-event of interest have no direct shape restrictions. This flexibility allows for more realistic modeling of multi-state phenomena such as disease progression, and offers a more sensitive approach to assessing the relative effects of covariates through different state transitions.

Markov models, which require exponential waiting times in states and independence of transition times from state destinations, are often used even when waiting times between states are known not to be exponentially distributed. For example, a multi-state model with exponential waiting times was used by Longini et al. (1989) to model AIDS progression, yet inspection of these data reveal that Weibull

waiting times are more appropriate (Butler and Huzurbazar, 1997). Flowgraphs, on the other hand, are semi-Markov; that is, waiting time distributions between states may be non-exponential, and are allowed to depend on the destination state. While other semi-Markov approaches have extreme difficulty handling censoring, incomplete data, or multiple returns to a previous state, flowgraph models incorporate these realistic features easily (Yau and Huzurbazar, 2002). Given a stochastic process with continuous time transitions between conditionally independent states, flowgraph models allow for most standard waiting time distributions to be used in modeling state transitions. Furthermore, flowgraphs provide a method for accessing any partial or total waiting time distribution of interest within the system. Flowgraph models are developed in terms of branch-specific transition probabilities and moment generating functions (MGFs), and saddlepoint approximations are then used to convert composite or overall MGFs to waiting time probability density functions (PDFs), cumulative distribution functions (CDFs), survival functions, or hazard functions of interest. Estimation may be performed in either a Bayesian or frequentist framework, depending on the resulting quantities and decision-making capabilities desired.

Through application of flowgraph models to disease progression in clinical trials, our end goal is better parametric modeling of times-to-event and treatment effects for the purpose of surrogate endpoint evaluation. Often, parametric models are desired in this setting, especially when one would like to model treatment effects and easily obtain predictive distributions for a patient population with dissimilar covariate values (e.g., age or stage of disease) that might be observed in a future trial where the validated surrogate is used as the primary endpoint.

Restricting consideration to simple parametric models for time-to-event endpoints causes specific issues to emerge within a surrogacy evaluation setting. Common waiting time distributions, including the exponential, Weibull, and lognormal

models, generally have monotonic or unimodal hazard functions unlikely to be observed in practice. Furthermore, competing risk distributions are inherent to popular putative surrogate endpoints such as disease-free survival and progression-free survival, where the times-to-event of interest are composed of disease recurrence or progression and death, whichever occurs first. However, hazard function shape assumptions and underlying competing risks distributions are very often ignored in the parametric approach to surrogacy evaluation of time-to-event endpoints, which could lead not only to erroneous treatment effect estimates, but also to poor assessments of surrogacy. Popular semiparametric and nonparametric approaches to time-to-event modeling often suffer from less precise treatment effect estimates than can be obtained from parametric approaches, as well as predictive distributions that are more difficult to obtain. The assumption of proportional hazards accompanying Cox models, for example, may additionally be unrealistic for a given application, and can generally be avoided through flowgraphs. Relaxation of parametric assumptions in flowgraph modeling has yet to be explored, due to an inability to express semiparametric models in terms of moment generating functions.

To address these issues, we propose moving the surrogate endpoint evaluation of time-to-event endpoints to the flowgraph setting, particularly in disease settings where a multi-state paradigm exists and parametric modeling needs to be made more realistic. Hazard function shapes are not explicitly restricted in flowgraph modeling; as a result, flexible and multi-modal hazards for overall waiting times of interest are not unusual to observe, even for relatively simple systems. Modeling of competing risks and recurrent events follow immediately from the inclusion of parallel branches and loops, respectively. For example, covariates such as treatment assignment may affect specific transition probabilities or branch waiting times within a flowgraph model for disease-free survival, and to a detailed extent that may be "washed out" by a simpler approach. Detailed knowledge of how covariates affect specific disease

state transitions, and not just the overall event times of interest, will certainly yield a more realistic analysis of surrogacy.

The remainder of the paper is organized as follows: in Section 4.2, we review existing flowgraph methods, including model construction, flowgraph algebra, saddlepoint approximation of required distribution functions, Bayesian estimation of parameters and predictive distributions, and incorporation of covariates. In Section 4.3, we develop flowgraph models for two clinical trials in colorectal cancer. Section 4.4 describes how the unique output of flowgraph models, including predictive densities of composite endpoints (such as disease-free survival) and posterior distributions of covariate effects, may then be used in surrogacy analyses involving time-to-event endpoints.

4.2 Review of Flowgraph Models and Methods

We begin by reviewing the basic construction of flowgraphs, general modeling of flowgraph components, and estimation methods for quantities of interest. Throughout, we adopt the notation and perspective of Huzurbazar (2005a).

4.2.1 Model Construction and Transmittances

Construction of flowgraph models is a flexible and logical process, readily informed by available knowledge of the real-world system under investigation. Like other multi-state modeling approaches, flowgraphs in their most basic form are made up of nodes and branches. However, in other graphical models, nodes usually represent random variables with branches representing relationships between these variables. In flowgraph models, nodes represent various states or events associated with a system, while branches represent transitions with waiting times in the previous state. Specifically, a *flowgraph* is a graphical representation of a stochastic system in which each directed branch is labeled with a *transmittance*, defined as the probability of following the branch times the moment generating function (MGF)

associated with the waiting time in the initiating branch state (Huzurbazar, 1999).

Consider for example the simple flowgraph presented in Figure 4.1. This system contains three states, where all individuals (or units) begin in the first state and then proceed either to state 2 with probability p_{12} or state 3 with probability $p_{13} = 1 - p_{12}$. In this case, both states 2 and 3 are terminal (absorbing) states, and the two possible paths from beginning to end of the total system may be represented as $1 \rightarrow 2$ and $2 \rightarrow 3$. In Figure 4.1, $M_{12}(s)$ and $M_{13}(s)$ are the MGFs associated with the waiting time distributions from state 1 to state 2, and state 1 to state 3, respectively. The product of the transition probability and moment generating function for a single branch is the transmittance; in this example, the two branch transmittances of the system are given by $p_{12}M_{12}(s)$ for transition $1 \rightarrow 2$ and $p_{13}M_{13}(s)$ for transition $1 \rightarrow 3$.

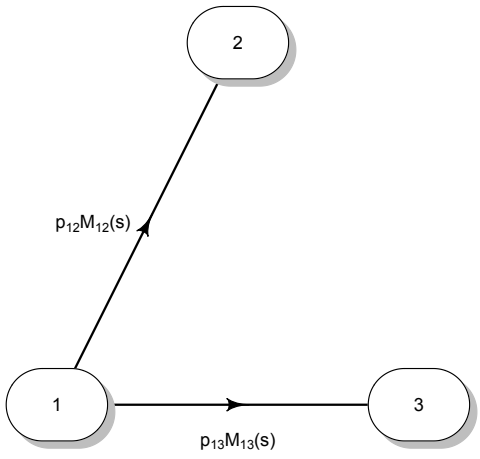

Figure 4.1. Simple flowgraph model with two parallel branches and absorbing states.

If Y_1 represents the waiting time from state 1 to state 2, and Y_2 similarly represents the waiting time from state 1 to state 3, the total flowgraph shown in Figure 4.1 describes the distribution of first passage, or the waiting time for the occurrence of $min\{Y_1, Y_2\}$. When Y_1 and Y_2 are assigned distribution functions, the transition probabilities become $p_{12} = P(Y_1 < Y_2)$ and $p_{13} = 1 - p_{12} = P(Y_2 < Y_1)$. It should be noted that $M_{12}(s)$ is not the MGF of Y_1, but rather the MGF of Y_1 *given* $Y_1 < Y_2$. In this way, flowgraphs readily handle conditional waiting time distributions based on competing risks (Huzurbazar, 1999).

In practice, basic structures other than parallel branches may appear within a given flowgraph; for example, a series structure where progression from state i to state j occurs with probability $p_{ij} = 1$, or a feedback loop where individuals remain in state i with probability $p_{ii} < 1$.

4.2.2 Solving Flowgraphs and Mason's Rule

After one has constructed a flowgraph to represent a real-world system as described in Section 4.2.1, the next step involves "solving" the flowgraph, or reducing the individual branch transmittances to equivalent transmittances for possible paths, thereby ultimately recovering the MGF of the entire system. In general, reduction to equivalent transmittances may be performed in a piecewise fashion, and the process is aided by a number of basic rules. In particular, reduced transmittances of branches in series are products, while reduced transmittances of branches in parallel are sums of the branch MGFs weighted by their associated transition probabilities. In the example accompanying Figure 4.1, the equivalent (reduced) transmittance for branches $1 \rightarrow 2$ and $1 \rightarrow 3$ is $T_E(s) = p_{12}M_{12}(s) + p_{13}M_{13}(s)$. Because this simple flowgraph consists only of three states with two parallel branches, $T_E(s)$ is also the overall MGF of the flowgraph corresponding to the waiting time distribution of $min\{Y_1, Y_2\}$.

Solving more complicated flowgraphs containing feedback loops and compositions of the preceding basic structures requires a more general formula as given in (4.1), which first depends on identification of each possible path and loop through the system. In this context, a *path* is defined as any possible sequence of nodes from input to output that does not pass through any intermediate node more than once. A *first-order loop* is defined as any closed path returning to the initiating node without passing through any other node more than once, while a jth-*order loop* consists of j non-touching first-order loops. The equivalent transmittance of a first-order loop is the product of the involved individual transmittances, while the equivalent transmittance of a jth-order loop is the product of the equivalent transmittances of the first-order loops it contains (Huzurbazar, 1999). In the example accompanying Figure 4.1, the paths are $1 \to 2$ and $1 \to 3$ with no intermediate nodes, and no loops are present.

With all paths and loops identified, Mason's rule (Mason, 1953) in its general form can be used to solve a flowgraph, or equivalently, to obtain the equivalent transmittance (MGF) of an entire flowgraph. It is given by

$$T_E(s) = \frac{\sum_i P_i(s)[1 + \sum_j (-1)^j L_j^i(s)]}{1 + \sum_j (-1)^j L_j(s)}, \tag{4.1}$$

where $P_i(s)$ is the transmittance for the ith path, $L_j(s)$ is the sum of the transmittances over the jth-order loops, and $L_j^i(s)$ is the sum of the transmittances over jth order loops sharing no common nodes with the ith path (Huzurbazar, 1999).

In the simple example presented in Figure 4.1, no loops are present, and the individual path transmittances for $1 \to 2$ and $1 \to 3$ are given by $P_1(s) = p_{12}M_{12}(s)$ and $P_2(s) = p_{13}M_{13}(s)$, respectively. Using Mason's rule, the overall transmittance (MGF) for this flowgraph is then given by $T_E(s) = T_E(s) = p_{12}M_{12}(s) + p_{13}M_{13}(s)$, as stated above.

In practice, distributional forms for each transition (i.e., exponential, Weibull) are chosen in advance, perhaps through exploratory analyses or model fit tests based

on available transition-specific data. Huzurbazar (2005a) further suggest visual inspection of "censored-data histograms" during the model selection phase, where such histograms are derived from Kaplan-Meier estimates and overlaid with a variety of fitted common waiting time distributions (see Ch. 6). Once general distributional forms have been chosen for each transition in a given flowgraph, the moment generating functions corresponding to these distributions are substituted for the $M_{\cdot}(s)$ terms in the overall MGF resulting from Mason's rule. Estimation of model parameters may then be performed through a Bayesian or classical approach.

4.2.3 Saddlepoint Approximation and Bayesian Estimation

With the total flowgraph equivalent transmittance (MGF) obtained as in Section 4.2.2, the end result desired (perhaps in a Bayesian clinical trial) is often a predictive distribution of future observables given current data. Other quantities of interest may be predictive hazard functions or survival functions of partial or overall waiting times from the flowgraph, or posterior distributions of individual or shared model parameters. All of these first require analytic or approximate numerical inversion of the flowgraph MGF. In what follows, we consider a Bayesian approach to estimation; classical approaches such as maximum likelihood estimation may also be used where possible.

When the MGF of the overall waiting time for a simple system is a convolution or mixture of exponential or gamma distributions, it may possible to invert algebraically via partial fraction expansion or numerical inversion. More complex models, or the presence of any other waiting time distributions such as Weibull or inverse Gaussian, will require alternative methods for approximate inversion. Huzurbazar (1999) suggest univariate saddlepoint approximation in such cases, which only requires the individual transition MGFs to be tractable (Daniels, 1954).

An algorithm for saddlepoint approximation of predictive densities and CDFs of interest for a given flowgraph was presented in Butler and Huzurbazar (2000). The authors note that their use of saddlepoint approximation differs from ordinary usage where $f(x|\theta)$ is approximated as a function of x with θ held fixed; instead, $f(x|\theta)$ is approximated as a function of θ with x fixed. Furthermore, predictive densities are often heavy-tailed with ill-defined moments, so that direct saddlepoint inversion of the marginal densities $f(x, z)$ and $f(x)$ is problematic. In what follows, Butler and Huzurbazar (2000) circumvent this issue by inverting conditional on θ and subsequently mixing over θ to remove dependence.

The predictive distribution of a waiting time Z with density $f(z|\theta)$ conditional on data x is given by

$$f(z|x) = \frac{f(x,z)}{f(x)} = \frac{E^\theta\{f(z|\theta)f(x|\theta)\}}{E^\theta\{f(x|\theta)\}}, \tag{4.2}$$

with associated CDF

$$F(z|x) = \frac{E^\theta\{F(z|\theta)f(x|\theta)\}}{E^\theta\{f(x|\theta)\}}. \tag{4.3}$$

Above, $f(x|\theta)$ is the data likelihood, and the expectation E^θ is with respect to the prior distribution $f(\theta)$. If the likelihood and resulting posterior distributions are intractable, saddlepoint approximations for the predictive density (4.2) and CDF (4.3) are jointly given by

$$\{\hat{f}(z|x), \hat{F}(z|x)\} = \frac{\hat{E}^\theta[\{\hat{f}(z|\theta), \hat{F}(z|\theta)\}\hat{f}(x|\theta)]}{\hat{E}^\theta\{\hat{f}(x|\theta)\}}, \tag{4.4}$$

where $\hat{E}^\theta$ denotes an approximate prior expectation. Computation of both $\hat{f}(z|x)$ and $\hat{F}(z|x)$ may be performed simultaneously, for example by simulation of $\theta_1, ..., \theta_m$ from a proper prior $f(\theta)$ to obtain

$$\{\hat{f}(z|x), \hat{F}(z|x)\} = \frac{\sum_{i=1}^m \{\hat{f}(z|\theta_i), \hat{F}(z|\theta_i)\}\hat{f}(x|\theta_i)}{\sum_{i=1}^m \hat{f}(x|\theta)}. \tag{4.5}$$

If the posterior distribution is tractable due to the likelihood, (4.5) can be simplified as

$$\{\hat{f}(z|x), \hat{F}(z|x)\} = m^{-1} \sum_{i=1}^{m} \{\hat{f}(z|\theta_i), \hat{F}(z|\theta_i)\}, \tag{4.6}$$

where $\theta_1, ..., \theta_m$ are draws from the posterior $f(\theta|x)$.

Whether (4.5) or (4.6) are used to obtain predictive distributions of interest, computation of saddlepoint approximations for $f(z|\theta)$ and/or $F(z|\theta)$ will likely be required. Suppose the moment generating function $M_Z(s|\theta)$ and cumulant generating function $K_Z(s|\theta) = \ln\{M_Z(s|\theta)\}$ are convergent on an open neighborhood of zero given by $s \in (a(\theta), b(\theta))$. Daniels (1954) gives a saddlepoint approximation for $f(z|\theta)$ as

$$\hat{f}(z|\theta) = \{2\pi K_Z''(\hat{s}|\theta)\}^{-1/2} \exp\{K_Z(\hat{s}|\theta) - \hat{s}z\}, \tag{4.7}$$

where $\hat{s} = \hat{s}(z, \theta)$ is the unique solution to the saddlepoint equation $K_Z'(\hat{s}|\theta) = z$ in $(a(\theta), b(\theta))$ when z is inside the convex support of $f(z|\theta)$. The accuracy of (4.7) in reproducing $f(z|\theta)$ is usually quite high over a wide range of values of z and θ; further details are provided in Butler and Huzurbazar (2000).

An approximation of $F(z|\theta)$ based on Luganni and Rice (1980) as described in Daniels (1987) is given by

$$\hat{F}(z|\theta) = \begin{cases} \Phi(\hat{w}) + \phi(\hat{w})(\hat{w}^{-1} - \hat{u}^{-1}), & z \neq E(Z|\theta), \\ 1/2 + K_Z'''(0|\theta)\{K_Z''(0|\theta)\}^{-3/2}/(6\sqrt{2\pi}), & z = E(Z|\theta), \end{cases} \tag{4.8}$$

where Φ and ϕ are the standard normal CDF and density, respectively, and

$$\hat{w} = \text{sgn}(\hat{s})\sqrt{2\{\hat{s}z - K_Z(\hat{s}|\theta)\}} \quad \text{and} \quad \hat{u} = \hat{s}\sqrt{K_Z''(\hat{s}|\theta)}$$

are implicit functions of z and θ according to $K_Z'(\hat{s}|\theta) = z$ (Butler and Huzurbazar, 2000).

Predictive distributions in (4.5) or (4.6) are computed over a grid of z-values, $z_1 < ... < z_n$, chosen to be sufficiently fine to capture all relevant detail of the

distribution. For each sampled value of θ from the prior, the saddlepoint equation (4.7) may be solved to obtain a draw from $\hat{f}(z|x)$. When the solution to (4.7) is not explicit in θ, a specialized algorithm such as that suggested in Butler and Huzurbazar (2000) may be required. When the posterior $f(\theta|x)$ is not in closed form, these authors propose renormalization of the predictive densities as

$$\tilde{f}(z_k|x) = \frac{\hat{f}(z_k|k)}{\sum_{i=1}^{n} \hat{f}(z_i|x)}, \quad k = 1, ..., n,$$

where $\hat{f}(\cdot|x)$ is given by (4.5) or (4.6). Otherwise, when the posterior is explicit, samples $\theta_1, ..., \theta_m$ from the posterior may be used to obtain the posterior expectation of individually renormalized densities as

$$\tilde{f}(z_k|x) = m^{-1} \sum_{i=1}^{m} \left\{ \frac{\hat{f}(z_k|\theta_i)}{\sum_{j=1}^{n} \hat{f}(z_j|\theta_i)} \right\}, \quad k = 1, ..., n.$$

A related algorithm for Bayesian prediction is described in Huzurbazar (2005b), with specific attention to the construction of likelihood components. With parametric models chosen for each state transition, the overall likelihood function consists of terms separable by these transitions, perhaps with censoring represented through survival functions. Given a choice of prior distribution for each model parameter (including transition probabilities), posterior distributions for each transition's parameter vector and transition probability may then be obtained via standard Markov chain Monte Carlo techniques such as Gibbs, Metropolis-Hastings, or slice sampling (see, e.g., Carlin and Louis (2009)). At this point, each random sample $\theta_1, ..., \theta_m$ from the joint posterior can be used to evaluate both the data likelihood $L(\theta_k|x)$ and the saddlepoint approximation $\hat{f}(z|\theta_k)$ corresponding to the overall MGF obtained from application of (4.1). These evaluations in turn, over the entire posterior sample, may be used to approximate the desired predictive density for a future observable Z by

$$f(z|x) = \frac{\int f(z|\theta) L(\theta|x) f(\theta) d\theta}{\int L(\theta|x) f(\theta) d\theta} \equiv E_{\theta|x}\{f(z|\theta)\}.$$

With estimated predictive densities or CDFs of interest at hand, other quantities of interest, such as hazard functions, survival functions, or quantiles may be easily obtained.

4.2.4 Extension: Inclusion of Covariates

Huzurbazar and Williams (2010) recently formalized the inclusion of covariates in the transition probabilities and waiting times of flowgraph models. In general, covariates are incorporated through branch model parameters in much the same way as in usual regression approaches to survival analysis. Covariates may also influence the branch transition probabilities, perhaps through standard logistic regression models. For example, in Figure 4.1, the transition probability p_{12} may depend on a single covariate x_1 through $\text{logit}(p_{12}) = \beta_0 + \beta_1 x_1$. If the $1 \to 2$ waiting time is further assigned an exponential distribution with parameter λ_{12}, the same covariate may influence the transition model, perhaps through $\lambda_{12} = \exp(\delta_0 + \delta_1 x_1)$. In some cases, a transition probability may also be written as a function of the associated and competing branch model parameters. Likelihood construction and Bayesian estimation then proceed as in Section 4.2.3, where censoring may be present. The MGF of the overall flowgraph model may still be obtained via (4.1), but now contains the additional complexity of observation-specific covariates nested within branch transmittances and probabilities. Huzurbazar and Williams (2010) note that prior selection is an important consideration in transitions having sparse data, particularly when additional regression coefficients must be estimated.

4.3 Flowgraph Models for Colon Cancer Trials

We now consider construction of flowgraphs for two recently completed trials in colorectal cancer. N0147 is a phase III, two-arm trial for adjuvant therapy in colon cancer (Albert et al., 2005). Patients in N0147 have stage III disease with completely resected tumors, and are randomized to modified FOLFOX6 (oxaliplatin

plus 5-FU/LV), a chemotherapeutic agent, with or without cetuximab, a biologic agent. The primary endpoint of interest in this trial is disease-free survival (DFS), which was recently validated as a surrogate endpoint for overall survival (Sargent et al., 2005). N0147 reached full accrual in November 2009, and patient follow-up is currently underway. So far, we know that individuals enrolled in this trial generally follow basic patterns of disease progression: patients start in an initial disease-free state (state 0) after surgical resection of their tumors, and may eventually experience recurrence (state 1), death (state 2), or recurrence followed by death. Additionally, we know that some patients return to a disease-free state after resection of recurred tumors. With this information, we construct the flowgraphs for overall survival and disease-free survival shown in the left-hand panels of Figure 4.2.

N9741, a phase III three-arm trial testing three therapies for advanced colorectal cancer, is our second trial to consider (Goldberg et al., 2004). Patients in N9741 had metastatic, non-resectable disease and were randomized to receive one of the following: IFL (irinotecan plus 5-FU/LV), FOLFOX (oxaliplatin plus 5-FU/LV), or irinotecan/oxaliplatin. The trial's primary results demonstrated that FOLFOX significantly increased overall survival when compared to the standard-of-care IFL regimen, and with fewer adverse events. Individuals enrolled in N9741 followed a similar disease progression structure to those in N0147, but with different events due to their advanced stage of disease. Beginning in a non-progressed disease state (0), patients may experience progressed disease (state 1) defined by significant tumor growth, death (state 2), or progressed disease followed by death. Similar to N0147, some patients may return to the initial non-progressed disease state if treatment is efficacious (defined by tumor shrinkage). However, this trial presents an additional modeling complexity: based on available follow-up, we notice a relatively small group of patients who, after starting in or returning to the non-progressed disease state (state 0), seem to remain there indefinitely. We loosely refer to this

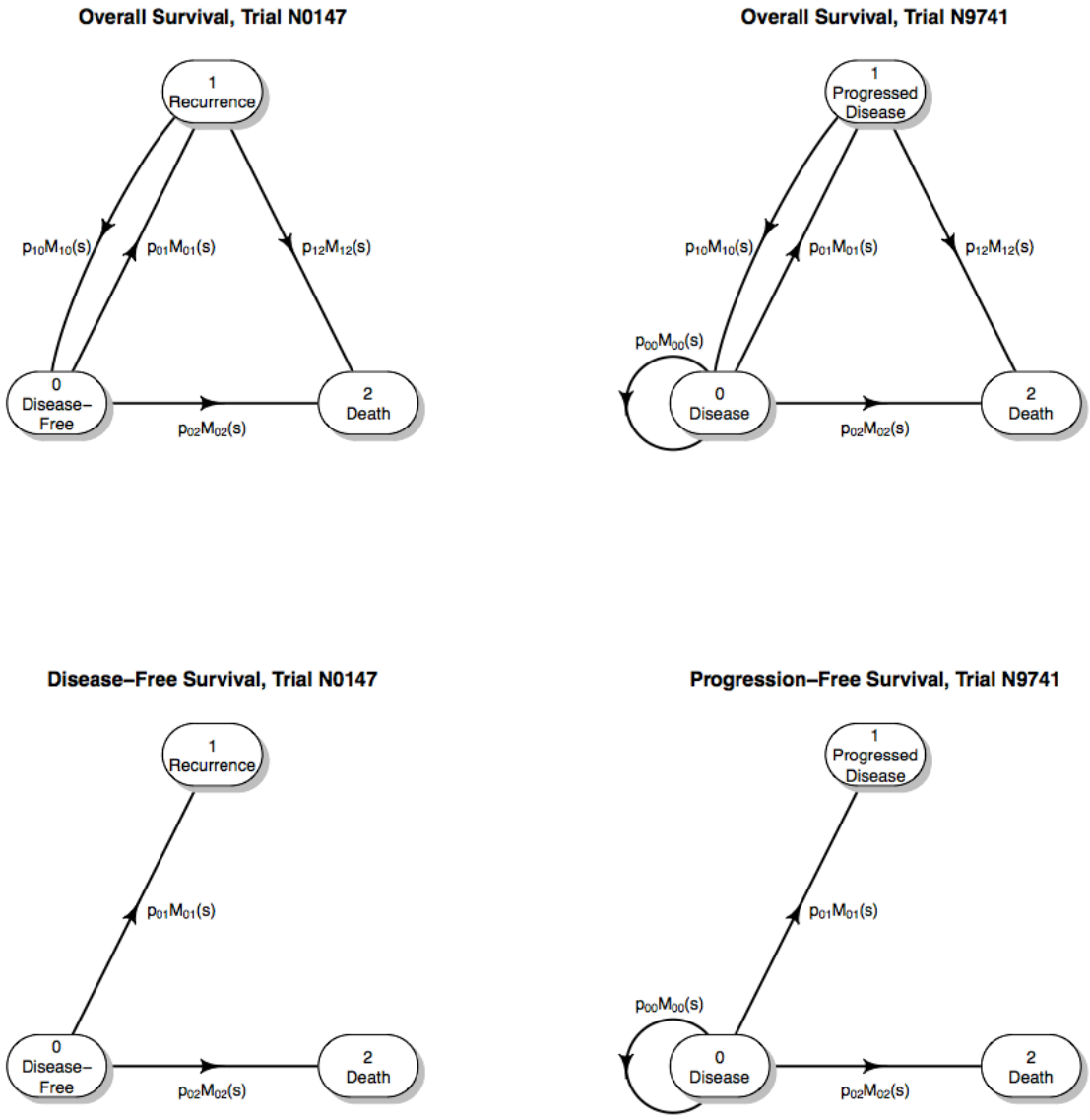

Figure 4.2: Flowgraph models for overall survival in trials N0147 and N9741, disease-free survival in trial N0147, and progression-free survival in trial N9741.

group of patients as "cured" patients, and conservatively modify the flowgraph to account for them. A direct feedback loop from state 0 to state 0 allows some of the patients to remain in state 0 with probability p_{00} and waiting time MGF $M_{00}(s)$. If a heavy-tailed distribution is chosen for $M_{00}(s)$, the waiting times of the cured patients may be modeled as indefinitely long compared to the other patients. The flowgraphs for overall survival and progression-free survival constructed for N9741 are shown in the right-hand panels of Figure 4.2.

4.3.1 Trial N0147: Reversible Illness-Death Model

The flowgraph for overall survival in trial N0147 happens to be a common multi-state model, sometimes called the "reversible illness-death model," used to describe cancer progression. Each branch of the model, $i \to j$, is labeled with its associated transmittance, or the probability of taking the branch (p_{ij}) times the MGF of the waiting time for that transition, $M_{ij}(s)$. To obey probabilistic laws, all transition probabilities corresponding to exits from a single state must sum to 1; in this case, we require $p_{01} + p_{02} = 1$ and $p_{10} + p_{12} = 1$. State 2, death, is an absorbing state; that is, exit from this state is not possible.

With the general form of the overall survival flowgraph identified as in the upper left panel of Figure 4.2, we identify two paths: a lower path $0 \to 2$ and an upper path $0 \to 1 \to 2$. The loop $0 \to 1 \to 0$ may initiate either path (perhaps repeatedly), but passage through this loop is not required. In either case, we can reduce the two series transmittances appearing in the $0 \to 1 \to 0$ loop to obtain an equivalent loop transmittance of $p_{01}M_{01}(s)p_{10}M_{10}(s)$. The lower path starting in state 0 either immediately returns to state 0 with this transmittance, or progresses to state 2 with transmittance $p_{02}M_{02}(s)$. Using an engineering-based balance equation approach or the fact that waiting time to exit (perhaps multiple passes through) a feedback loop follows a geometric distribution (see Huzurbazar (2005a), Ex. 2.14,

pp. 30-32), we solve for the total lower path transmittance to obtain

$$T_{lower}(s) = p_{02}\left[\frac{M_{02}(s)}{1 - p_{01}p_{10}M_{01}(s)M_{10}(s)}\right]. \tag{4.9}$$

For the upper path, we may choose instead to reduce the loop transmittance to state 1 (rather than state 0 as above), so that the upper path begins in state 0 and progresses to state 1, where it may loop back to state 1 before proceeding to state 2. The $1 \to 1$ and $1 \to 2$ transmittances may then be reduced to a single $1 \to 2$ transmittance using the same balance equation or geometric approach referenced in the construction of (4.9). Finally, the $0 \to 1$ and $1 \to 2$ transmittances in series reduce to the overall equivalent transmittance for the upper path:

$$T_{upper}(s) = p_{01}\left[\frac{p_{12}M_{01}(s)M_{12}(s)}{1 - p_{01}p_{10}M_{01}(s)M_{10}(s)}\right]. \tag{4.10}$$

The resulting upper and lower path transmittances given in (4.9) and (4.10) may now be viewed as parallel branches to obtain the overall survival MGF in trial N0147:

$$M(s) = \frac{p_{01}p_{12}M_{01}(s)M_{12}(s) + p_{02}M_{02}(s)}{1 - p_{01}p_{10}M_{01}(s)M_{10}(s)}. \tag{4.11}$$

We note that the flowgraph above could have been solved using Mason's rule (4.1) with the same result. With specific distributional forms chosen for each branch-specific waiting time MGF, the methods discussed in Section 4.2 may be used to obtain predictive distributions for death times or other quantities of interest for the N0147 trial, where both right-censoring and patient covariates are certain to be present.

In the context of surrogacy evaluation, overall survival (OS) is a likely clinical endpoint for trial N0147, while disease-free survival (DFS) might be a desirable potential surrogate to replace OS. Thus, in addition to the full flowgraph developed for OS above, we may construct a reduced flowgraph for overall time to DFS, defined as the earlier of disease recurrence or death. This flowgraph, shown in the lower left panel of Figure 4.2, consists of the same system without the $1 \to 2$ and $1 \to 0$

branches, where state 1 becomes a second possible absorbing state. Because the first-order loop present for OS no longer exists in this framework, the flowgraph for DFS simply consists of two parallel branches, with overall MGF from Mason's rule given by $M(s) = p_{01}M_{01}(s) + p_{02}M_{02}(s)$. Estimation of the model parameters and regression coefficients associated with each flowgraph will yield Bayesian predictive distributions and treatment effect vectors for DFS and OS, which may then be used to evaluate surrogacy (see Section 4.4).

4.3.2 Trial N9741: Illness-Death Model with a Possibly Prolonged Initial State

The flowgraph for overall survival in trial N9741, shown in the upper right panel of Figure 4.2, is similar to the reversible illness-death flowgraph for overall survival in trial N0147 discussed in Section 4.3.1. However, a direct feedback loop $0 \to 0$ adds additional modeling complexity. We solve this flowgraph using Mason's rule (4.1) as presented in Section 4.2.2. Recall when using (4.1), a path is defined as any possible sequence of nodes from input to output that does not pass through any intermediate states more than once. According to this definition, there are two paths, $0 \to 1 \to 2$ and $0 \to 2$. The transmittances corresponding to these paths are $P_1(s) = p_{01}M_{01}(s)p_{12}M_{12}(s)$ and $P_2(s) = p_{02}M_{02}(s)$, respectively. Furthermore, we have two first-order loops given by $0 \to 0$ and $0 \to 1 \to 0$. The sum of the transmittances over these first-order loops is $L_1(s) = p_{00}M_{00}(s) + p_{01}M_{01}(s)p_{10}M_{10}(s)$. There are no higher-ordered loops present in the flowgraph, and neither first-ordered loop avoids sharing common nodes with each path. Thus, $L_j^i = 0$ for all i, j, and the bracketed quantity in the numerator of (4.1) is simply 1. The overall survival MGF of the flowgraph for trial N9741 is then given by Mason's rule as

$$M(s) = \frac{p_{01}p_{12}M_{01}(s)M_{12}(s) + p_{02}M_{02}(s)}{1 - p_{00}M_{00}(s) - p_{01}M_{01}(s)p_{10}M_{10}(s)}. \tag{4.12}$$

When choosing the branch-specific waiting time distributional forms for overall survival in trial $N9741$, we would give special attention to the choice of distribution

for the direct feedback loop $M_{00}(s)$. Recall that we incorporated this loop into the N9741 flowgraph to handle patients who seem to remain in state 0 for a prolonged period, perhaps until they are right-censored in this state after some period of follow-up. For this reason, we would intentionally choose a heavy-tailed distribution, such as inverse Gaussian, to ensure lengthy stays in state 0 are possible for some patients. Other branch-specific distributional forms may be individually selected through exploratory analyses or visual inspection of censored-data histograms (Huzurbazar, 2005a). Ultimately, the saddlepoint approximation and Bayesian methods of Section 4.2 may be used to obtain estimates and predictions of interest for the N9741 trial, where right-censoring and covariates will accompany consideration of the special group of "cured" patients.

In trial N9741, we may wish to investigate progression-free survival (PFS) as a candidate surrogate for overall survival (OS), where PFS is defined as the earlier of disease progression or death. In this setting, the full flowgraph for OS presented above may be reduced to a separate flowgraph for PFS, where the $1 \to 0$ and $1 \to 2$ transitions are removed and state 1 is now a second absorbing state. This flowgraph, shown in the lower right panel of Figure 4.2, retains the $0 \to 0$ first-order loop in addition to two parallel branches given by $0 \to 1$ and $0 \to 2$. Using Mason's rule, the overall MGF for PFS is then given by $M(s) = [p_{01}M_{01}(s) + p_{02}M_{02}(s)]/p_{00}M_{00}(s)$. Estimation of parameters associated with each flowgraph will yield quantities useful for the surrogacy evaluation of PFS for OS (see Section 4.4).

4.4 Application to Surrogacy Evaluation

Ultimately, a surrogate endpoint S for a true endpoint T must be validated at two levels: the within-trial or patient level, and the across-trial or meta-analytic level. Common surrogacy measures for time-to-event endpoints include copula R^2_{indiv} at the patient level and R^2_{trial} at the trial level, both discussed within a frequentist

framework by Burzykowski et al. (2001). However, the quantities obtained from flowgraph modeling are distinctly different from observed endpoints S, T and estimated treatment effect pairs $\hat{\beta}_S, \hat{\beta}_T$ used to assess surrogacy at the patient and trial levels, respectively, within existing evaluative frameworks in the surrogate endpoints literature.

Using flowgraph models, patient-level surrogacy might now be evaluated based on the joint predictive distribution of S and T, $f(z_s, z_t|x)$, obtained through Bayesian estimation of the flowgraphs that describe S and T. If the predictive distributions for S and T are highly associated within patient groups such as treatment arms, S may be a good surrogate for T at the patient level. With data available from multiple similar clinical trials indexed by $i \in 1, ..., N$, trial-level surrogacy could be assessed through hierarchical modeling of the joint treatment effect *vectors* on S and T, denoted $(\beta_{S,i}, \beta_{T,i})$, where the dimension of each vector follows from the number of branch transmittances and transition probabilities containing regression components for treatment effect in each of the two flowgraphs for S and T. These vectors, estimated from each pair of flowgraphs representing S and T across trials, could then be studied for strongly-related and consistent S, T relationships across trials. Flowgraphs become important to this goal through improved modeling of treatment effect covariates, with effects allowed to be specific to each state transition rather than assumed fixed over an entire flowgraph or partial flowgraph. For example, the effect of treatment on the transition from a disease-free state to recurrence may be vastly different than the effect of the same treatment on the transition from recurrence to death. Using this added information in surrogacy analyses would promote more sophisticated identification of candidate surrogate endpoints.

CHAPTER FIVE

Conclusion

5.1 Bayesian Evaluation of Surrogate Time-to-Event Endpoints

In Chapter 2, we have demonstrated that using a straightforward Bayesian approach with vague prior distributions generally enhances the stability and performance of both unadjusted and adjusted trial-level surrogacy measures, with Bayesian R^2_{adj} often offering the greatest advantages. Caution undoubtedly must be used, however, as estimates of R^2_{adj} may be biased high in meta-analyses with too few trials, or when many data characteristics (trial size, censoring rate, range of treatment effects) are simultaneously moved away from ideal levels. In the secondary simulations, this bias is not as extreme as that demonstrated by the unadjusted measure in the opposite direction. Still, we remain concerned that R^2_{adj} as originally formulated was not able to distinguish the likely vast difference in surrogacy of TTR for OS between the old and new ACCENT trials. Clearly, in consideration of the relationship between treatment effects on OS and treatment effects on TTR across trials, there are settings in which the adjusted model attributes too much of the deviation from perfect surrogacy to estimation error, rather than to underlying imperfect surrogacy. Furthermore, while this dissertation only presents results based on vague priors for unknown parameters in each Bayesian model, consideration of more informative priors on quantities within the adjusted model did not greatly influence or improve these results.

Even though we have only examined error-adjusted, trial-level surrogacy measures in the context of time-to-event endpoints, the lessons learned here may be beneficial for evaluations involving other endpoint types. As meta-analytic assessments of surrogate endpoints continue to surpass single-trial assessments in both

popularity and practicality (Burzykowski, 2008), we expect Bayesian modeling to play an expanding role in facilitating the implementation of increasingly complex evaluative approaches. For example, Abrahantes et al. (2004) performed a Bayesian simulation study to assess "coarse" to "fine" hierarchical structures for evaluation of normally distributed surrogate endpoints in a meta-analytic setting, accounting for the natural hierarchy and levels of information shared by patients within centers and centers within trials.

While recent papers on surrogate endpoint evaluation and validation still occasionally include discussions regarding the extent to which the Prentice criteria are satisfied by newer methods, some authors have noted how the Prentice criteria seem to imply that meta-analyses are *required* to prove surrogacy (Molenberghs et al., 2002). Another strong argument for meta-analytic evaluation of surrogate endpoints is provided by De Gruttola et al. (2001), who specifically promote the construction of databases of clinical trials on which patient-level as well as trial-level surrogacy analyses can be conducted. Efforts to compile such databases would be motivated by the fact that studies designed specifically to evaluate surrogates are generally implausible (Lesko and Atkinson, 2001), and *collections* of trials would allow for in-depth evaluation of potential surrogates across similar studies and within specific diseases. Certainly the ACCENT database considered here serves as an important model for other disease types, from the deliberate collection of similar trials to rigorous evaluation of candidate endpoints and ultimately, regulatory approval based on validated surrogates in future trials.

More sophisticated approaches to surrogacy evaluation than those considered here certainly do exist, with many developed specifically for broad application across endpoint types. Alonso et al. (2004) introduced the likelihood reduction factor (LRF) for general endpoints, to replace R^2-type measures specific to normal, survival, or discrete endpoints at the patient level. Later, Alonso and Molenberghs (2007) proposed

an information-theoretic approach to surrogate endpoint evaluation at both individual and trial levels within a meta-analytic setting, designed to accommodate general endpoint types. Bringing the evaluation of surrogacy into another realm, Frangakis and Rubin (2002) provided definitions of *principal surrogates* and *statistical surrogates*, based on the theory of causal inference. Such advances are mathematically elegant and offer certain theoretical advantages; in practice, however, the candidacy of specific types of endpoints, not general endpoints, are of interest. More importantly, it remains to be seen whether evaluative approaches for surrogate endpoints based on information theory and causal inference could ever become widely accessible to physicians, clinicians, and statisticians alike–an important goal for which the R^2-type measures have already demonstrated success (Shi and Sargent, 2009).

In light of more complex approaches to surrogacy evaluation, Fleming (2005a,b) encourages more broad, exploratory approaches similar to Sargent et al. (2005), to include a panel of intuitive meta-analytic measures, graphical checks, and measurement of concordance across trials for establishing the association between treatment effects on the true and candidate surrogate endpoints. Others have emphasized that a surrogate endpoint could never be validated by a single method (Green et al., 2008); rather, a candidate endpoint should demonstrate consistently high surrogacy across a range of measures. Moreover, statistical evaluation of surrogacy should complement the intuition of clinicians, those most familiar with the disease being treated and the mechanisms of action by which the intervention being studied may affect the surrogate and clinical endpoints of interest.

We have shown that R^2_{adj}, originally introduced to accommodate error in estimation of treatment effects in meta-analyses of survival endpoints, need not be abandoned due to its usual unavailability within the maximum likelihood paradigm. This theoretically more realistic measure of trial-level surrogacy is easily available within the Bayesian framework, and in our simulations, generally outperforms popularly-

used unadjusted classical measures in terms of bias, MSE, and coverage to a greater degree as trial characteristics become less ideal. Not surprisingly, instability that exists for all meta-analytic surrogacy measures in analyses based on too few trials or trials of low effective sample size continues to impact R^2_{adj}, but generally to a lesser degree than R^2_{un}.

In the ACCENT meta-analyses, highly precise estimation of surrogacy, especially when adjusting for error in estimation of the treatment effects, is certainly inhibited by rather small numbers of trials and other data characteristics. We must acknowledge, however, that the true surrogacy of TTR for OS is unknown–rather than fixed–in these examples. It could be that point estimates of Bayesian R^2_{adj} are optimistically high for each of the ACCENT analyses, but because of high posterior variability, we would correctly avoid putting too much confidence in TTR as a surrogate for OS. Although questions and concerns remain regarding widespread use of R^2_{adj}, a Bayesian approach makes it *possible* to obtain such error-adjusted measures for consideration. Furthermore, the extended method of Section 2.5 offers an attractive alternative – one we hope will be the subject of future investigation. As such, we strongly encourage exploration and inclusion of Bayesian measures in future meta-analyses for potential surrogate endpoints, as the advantages discovered here could yield similar insight for other endpoint types and disease settings.

5.2 Bayesian Adaptive Trial for a Newly Validated Surrogate

Although they have been used to establish efficacy and gain regulatory approval in practice, the very idea of surrogate endpoints continues to incite trepidation, from conflicting ideas regarding validation to concern regarding their ultimate implementation. In Chapter 3, we proposed a novel trial design that allows a newly-validated surrogate endpoint to play its intended role as the primary endpoint for determining the effect of an experimental treatment, while adaptively checking its

performance for consistency with historical data used for its validation. Understanding that practical concerns are likely to remain even after a surrogate has been deemed 'valid' for future use, our design quantifies knowledge and uncertainty from the validation stage and uses this information to assess the surrogate's performance during a new trial, while retaining the other advantages of Bayesian adaptivity. While we demonstrated our design in the context of survival endpoints, other types of endpoints and treatment effect measures could easily be used, with only minor modifications to our algorithm.

In both our simulation studies and application to adjuvant therapy trials in colon cancer, substantial savings in trial length and sample size were observed, while incorrect conclusions regarding surrogacy and effect of treatment were generally avoided (and avoided early). With patients, clinicians, and sponsors alike hoping to see beneficial new therapies approved and available for use as quickly as is reasonably possible, this design offers promise for those diseases where validated surrogate endpoints already exist or will exist in the future.

Finally, a concern that appears in the literature regarding hypothetical trials using surrogate endpoints is that shortened patient follow-up may be inadequate to detect rare or delayed adverse events (Wittes et al., 1989). While this point is certainly valid, it is also true for any trial that is designed to stop early or decrease in size based on available information. In general, we believe this risk is inherent to most (including classical) clinical trial designs, and pales in comparison to the many tangible and ethical advantages of adaptive design.

5.3 Flowgraph Modeling for Surrogacy Evaluation

In Chapter 4, we discussed the application of flowgraph modeling to cancer trials, with the goal of an improved and flexible parametric approach useful for evaluating surrogate endpoints in a more realistic manner. While we encourage the use

Bayesian methods, which yield predictive distributions of observables and posterior distributions on parameters of interest (rather than point estimates and approximate variances), a classical approach to flowgraph estimation is also possible. Yau and Huzurbazar (2002) offer a step-by-step treatment from a frequentist framework, including construction of likelihood components for complete and missing data, construction of the associated estimated total MGF, and numerical integration to obtain estimates of the total waiting time CDF, pdf, and hazard function. This paper includes handling of censoring and incomplete data (unknown transition times), but does not discuss inclusion of covariates. A drawback of this non-Bayesian approach is that variability of stage-specific parameter estimates is ignored, and thus parameter uncertainty does not carry forward in estimation of the overall model.

Huzurbazar and Williams (2010) discuss possible relaxation of the semi-Markov assumption to handle dependencies among subsequent waiting time distributions. When successive waiting times are not conditionally independent, the semi-Markov assumption no longer holds. Huzurbazar and Williams (2010) discuss how to alter likelihood functions and MGFs for this case, but warn that using time-dependent covariates and time-dependent transition probabilities in flowgraphs is especially challenging, quickly becoming impossible for complex models. The same paper also discusses handling of special types of recurrent events, and offers a perspective on Bayesian model averaging over predictive distributions resulting from flowgraphs. Huzurbazar and Williams (2010) further note that saddlepoint approximation may oversmooth some densities; in this case, Fourier transforms may be considered instead. Estimation of flowgraphs with incomplete data, or unobserved transition times that appear to "skip" states, is discussed in Huzurbazar (2000).

BIBLIOGRAPHY

Abrahantes, J. C., Molenberghs, G., Burzykowski, T., Shkedy, Z., Alonso, A., and Renard, D. (2004), "Choice of Units of Analysis and Modeling Strategies in Multilevel Hierarchical Models," *Computational Statistics and Data Analysis*, 47, 537–563.

Alberts, S. R., Sinicrope, F. A., and Grothey, A. (2005), "N0147: A Randomized Phase III Trial of Oxaliplatin Plus 5-Fluorouracil/Leucovorin With or Without Cetuximab After Curative Resection of Stage III Colon Cancer," *Clinical Colorectal Cancer*, 5, 211–213.

Alonso, A., and Molenberghs, G. (2007), "Surrogate Marker Evaluation From an Information Theory Perspective," *Biometrics*, 63, 180–186.

Alonso, A., Molenberghs, G., Burzykowski, T., Renard, D., Geys, H., Shkedy, Z., Tibaldi, F., Abrahantes, J. C., and Buyse, M. (2004), "Prentice's Approach and the Meta-Analytic Paradigm: A Reflection on the Role of Statistics in the Evaluation of Surrogate Endpoints," *Biometrics*, 60, 724–728.

Berry, S. M., Carlin, B. P., Lee, J. J., and Müller, P. (2010), *Bayesian Adaptive Methods for Clinical Trials*, New York: CRC Press.

Burzykowski, T. (2008), "Surrogate Endpoints: Wishful Thinking or Reality?" *Statistical Methods in Medical Research*, 17, 463–466.

Burzykowski, T., and Abrahantes, J. C. (2005), "Validation in the Case of Two Failure-Time Endpoints," in *The Evaluation of Surrogate Endpoints*, Burzykowski, T., Molenberghs, G., and Buyse, M. (eds.), New York, NY: Springer, 163–194.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., and Renard, D. (2001), "Validation of Surrogate End Points in Multiple Randomized Clinical Trials With Failure Time End Points," *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 50, 405–422.

Butler, R. W. and Huzurbazar, A. V. (1997), "Stochastic Network Models for Survival Analysis," *Journal of the American Statistical Association*, 92, 246–257.

—— (2000), "Bayesian Prediction of Waiting Times in Stochastic Models," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 28, 311–325.

Buyse, M., and Molenberghs, G. (1998), "Criteria for the Validation of Surrogate Endpoints in Randomized Experiments," *Biometrics*, 54, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000), "The Validation of Surrogate Endpoints in Meta-Analyses of Randomized Experiments," *Biostatistics*, 1, 49–67.

Carlin, B. P., and Louis, T. A. (2009), *Bayesian Methods for Data Analysis* (3rd ed.), New York: CRC Press.

Clayton, D. G. (1978), "A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence," *Biometrika*, 65, 141–151.

Cowles, M. K. (2004), "Evaluating Surrogate Endpoints for Clinical Trials: A Bayesian Approach," Technical Report, University of Iowa, Department of Statistics and Actuarial Science.

Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, Series B (Methodological)*, 34, 187–220.

Daniels, H. (1954), "Saddlepoint Approximations in Statistics," *Annals of Mathematical Statistics*, 25, 631–650.

—— (1987), "Tail Probability Approximations," *International Statistical Review*, 55, 37–48.

Daniels, M. J., and Hughes, M. D. (1997), "Meta-Analysis for the Evaluation of Potential Surrogate Markers," *Statistics in Medicine*, 16, 1965–1982.

De Gruttola, V., Clax, P., DeMets, D. L., Downing, G. J., Ellenberg, S. S., Friedman, L., Gail, M. H., Prentice, R. L., Wittes, J., and Zeger, S. L. (2001), "Considerations in the Evaluation of Surrogate Endpoints in Clinical Trials: Summary of a National Institutes of Health Workshop," *Controlled Clinical Trials*, 22, 485–502.

Dorf, R. (1989), *Modern Control Systems*, 5th ed, London: Addison-Wesley.

Fleming, T. R. (1994), "Surrogate Markers in AIDS and Cancer Trials," *Statistics in Medicine*, 13, 1423–1435.

—— (2005a), "Surrogate Endpoints and FDA's Accelerated Approval Process," *Health Affairs*, 24, 67–78.

—— (2005b), "Objective Response Rate as a Surrogate Endpoint: A Commentary," *Journal of Clinical Oncology*, 23, 4845–4846.

Frangakis, C. E., and Rubin, D. B. (2002), "Principal Stratification and Causal Inference," *Biometrics*, 58, 21–29.

Genest, C., and MacKay, J. (1986), "The Joy of Copulas: Bivariate Distributions with Uniform Marginals," *The American Statistician*, 40, 280–283.

Goldberg, R. M., Sargent, D. J., Morton, R. F., Fuchs, C. S., Ramanathan, R. K., Williamson, S. K., Findlay, B. P., Pitot, H. C., and Alberts, S. R. (2004), "A Randomized Controlled Trial of Fluorouracil Plus Leucovorin, Irinotecan, and Oxaliplatin Combinations in Patients With Previously Untreated Metastatic Colorectal Cancer," *Journal of Clinical Oncology*, 22, 23–30.

Green, E. M., Yothers, G., and Sargent, D. J. (2008), "Surrogate Endpoint Validation: Statistical Elegance Versus Clinical Relevance," *Statistical Methods in Medical Research*, 17, 477–480.

Hotelling, H. (1953), "New Light on the Correlation Coefficient and its Transforms," *Journal of the Royal Statistical Society, Series B*, 15, 193–232.

van Houwelingen, H. C., Arends, L. R., and Stijnen, T. (2002), "Advanced Methods in Meta Analysis: Multivariate Approach and Meta-Regression," *Statistics in Medicine*, 21, 589–624.

Hughes, M. D., De Gruttola, V., and Welles, S. (1995), "Evaluating Surrogate Markers," *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 10, S1–S8.

Huzurbazar, A. V. (1999), "Flowgraph Models for Generalized Phase Type Distributions Having Non-Exponential Waiting Times," *Scandinavian Journal of Statistics*, 26, 145–157.

—— (2000), "Modeling and Analysis of Engineering Systems Data Using Flowgraph Models," *Technometrics*, 42, 300–306.

—— (2005a), *Flowgraph Models for Multistate Time-to-Event Data*, Hoboken: John Wiley & Sons.

—— (2005b), "Flowgraph Models: A Bayesian Case Study in Construction Engineering," *Journal of Statistical Planning and Inference*, 129, 181–193.

Huzurbazar, S. and Huzurbazar, A. V. (1999), "Survival and Hazard Functions for Progressive Diseases Using Saddlepoint Approximations," *Biometrics*, 55, 198–203.

Huzurbazar, S. and Williams, B. J. (2010), "Incorporating Covariates in Flowgraph Models: Applications to Recurrent Event Data," *Technometrics*, 52, 198–208.

Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, 79–86.

Lesko, L. J., and Atkinson, A. J. Jr. (2001), "Use of Biomarkers and Surrogate Endpoints in Drug Development and Regulatory Decision Making: Criteria, Validation, Strategies," *Annual Review of Pharmacology and Toxicology*, 41, 347–366.

Longini, I., Clark, W., Byers, R., Ward, J., Darrow, W., Lemp, G., and Hethcote, H. (1989), "Statistical Analysis of the Stages of HIV Infection Using a Markov Model," *Statistics in Medicine*, 8, 831–843.

Luganni, R. and Rice, S. (1980), "Saddlepoint Approximations for the Distribution of the Sum of Independent Random Variables," *Advances in Applied Probability*, 12, 475–490.

Mason, S. J. (1953), "Feedback Theory – Some Properties of Signal Flow Graphs," *Proceedings of the Institute of Radio Engineers*, 41, 1144–1156.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T., and Alonso, A. (2002), "Statistical Challenges in the Evaluation of Surrogate Endpoints in Randomized Trials," *Controlled Clinical Trials*, 23, 607–625.

O'Brien, P. C., and Fleming, T. R. (1979), "A Multiple Testing Procedure for Clinical Trials," *Biometrics*, 35, 549–556.

Prentice, R. L. (1989), "Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria," *Statistics in Medicine*, 8, 431–440.

Sargent, D. J., Patiyil, S., Yothers, G., Haller, D. G., Gray, R., Benedetti, J., Buyse, M., Labianca, R., Seitz, J., O'Callaghan, C. J., Francini, G., Grothey, A., O'Connell, M., Catalano, P. J., Kerr, D., Green, E., Wieand, H. S., Goldberg, R. M., and de Gramont, A., ACCENT Group (2007), "End Points for Colon Cancer Adjuvant Trials: Observations and Recommendations Based on Individual Patient Data From 20,898 Patients Enrolled Onto 18 Randomized Trials From the ACCENT Group," *Journal of Clinical Oncology*, 25, 4569–4574.

Sargent, D. J., Shi, Q., Yothers, G., Van Cutsem, E., Cassidy, J., Saltz, L., Wolmark, N., Bot, B., Grothey, A., Buyse, M., and de Gramont, A., ACCENT Group (2011), "Two or Three Year Disease Free Survival as a Primary Endpoint in Stage III Adjuvant Colon Cancer Trials with Fluoropyrimidines With or Without Oxaliplatin or Irinotecan: Data From 12,676 Patients from MOSAIC, X-ACT, PETACC-3, C-06, C-07, and C89803," *European Journal of Cancer*, in press.

Sargent, D. J., Wieand, H. S., Haller, D. G., Gray, R., Benedetti, J. K., Buyse, M., Labianca, R., Seitz, J. F., O'Callaghan, C. J., Francini, G., Grothey, A., O'Connell, M., Catalano, P. J., Blanke, C. D., Kerr, D., Green, E., Wolmark, N., Andre, T., Goldberg, R. M., and de Gramont, A. D. (2005), "Disease-Free Survival Versus Overall Survival as a Primary End Point for Adjuvant Colon Cancer Studies: Individual Patient Data From 20,898 Patients on 18 Randomized Trials," *Journal of Clinical Oncology*, 23, 8664–8670.

Shi, Q., and Sargent, D. J. (2009), "Meta-Analysis for the Evaluation of Surrogate Endpoints in Cancer Clinical Trials," *International Journal of Clinical Oncology*, 14, 102–111.

Shih, J., and Louis, T. (1995), "Inferences on the Association Parameter in Copula Models for Bivariate Survival Data," *Biometrics*, 51, 1384–1399.

Tibaldi, F. S., Abrahantes, J. C., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003), "Simplified Hierarchical Linear Models for the Evaluation of Surrogate Endpoints," *Journal of Statistical Computation and Simulation*, 73, 643–658.

Whitehouse, G. (1983), "Flowgraph Analysis," in *Encyclopedia of Statistical Science* (Vol. 3.), eds. S. Kotz and N. Johnson, New York: Wiley.

Wittes, J., Lakatos, E., and Probstfield, J. (1989), "Surrogate Endpoints in Clinical Trials: Cardiovascular Diseases," *Statistics in Medicine*, 8, 415–425.

Yau, C. L. and Huzurbazar, A. V. (2002), "Analysis of Censored and Incomplete Survival Data Using Flowgraph Models," *Statistics in Medicine*, 21, 3727–3743.

Yothers, G. (2007), "Toward Progression-Free Survival as a Primary End Point in Advanced Colorectal Cancer," *Journal of Clinical Oncology*, 25, 5153–5154.